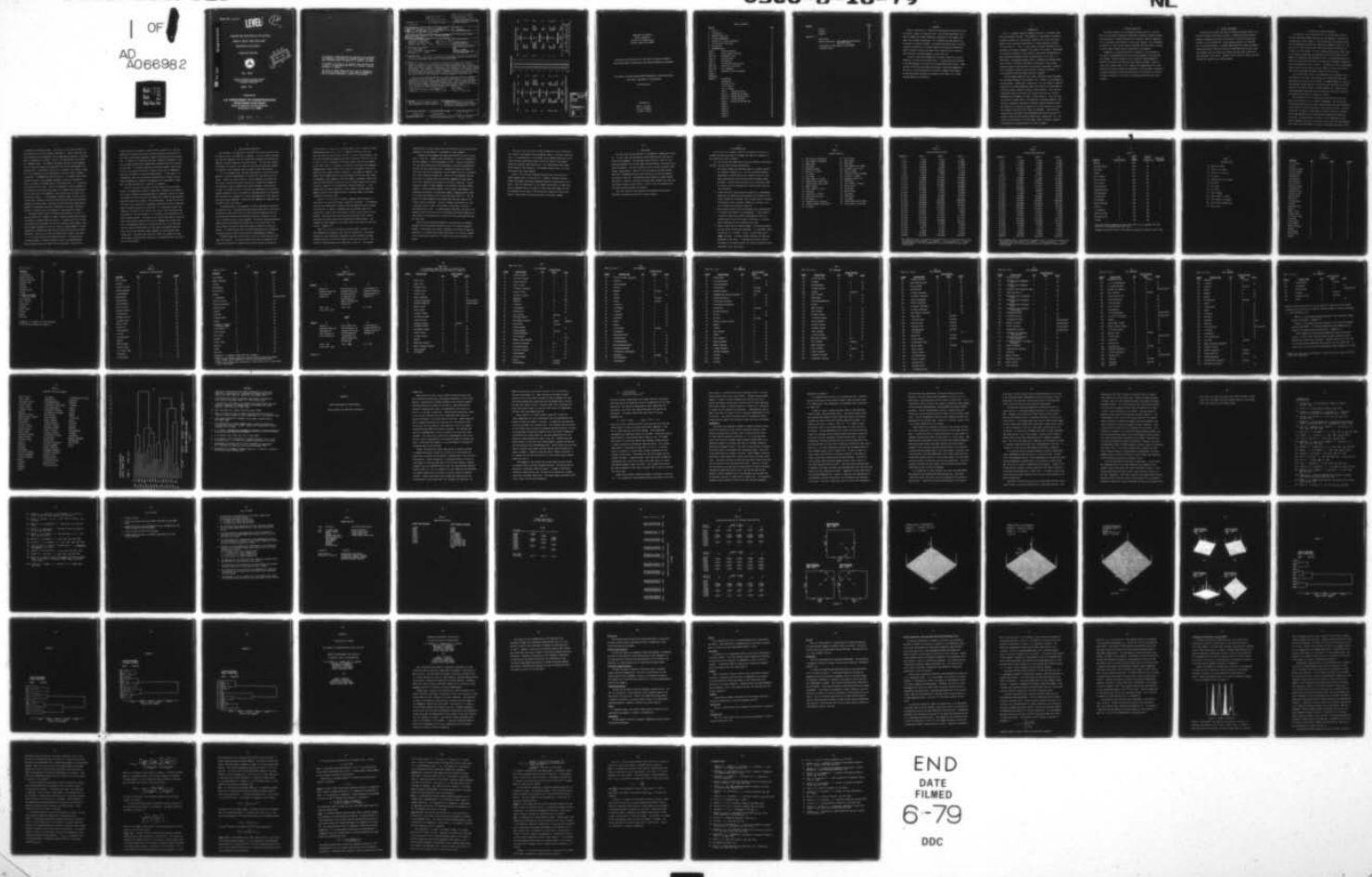
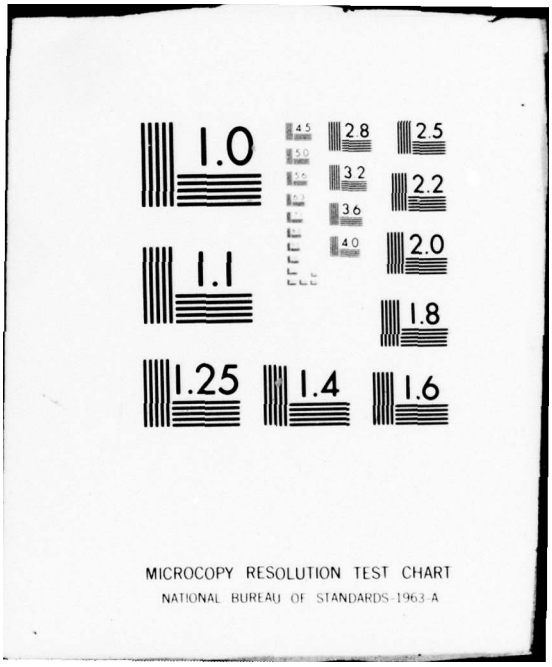
AD-A066 982 RHODE ISLAND UNIV KINGSTON DEPT OF CHEMISTRY F/G 7/1
EVALUATION AND CLASSIFICATION OF THE ELECTRICAL HAZARD OF CHEMI--ETC(U)
JAN 79 J L FASCHING, E W STROMBERG DOT-CG-44160-A
UNCLASSIFIED USCG-D-16-79 NL



END
DATE FILMED
6-79
DDC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

Report No. CG-D-16-79

LEVEL II

12

EVALUATION AND CLASSIFICATION OF THE ELECTRICAL HAZARD OF CHEMICAL VAPORS DURING WATER TRANSPORTATION USING PATTERN RECOGNITION TECHNIQUES

FINAL REPORT

JANUARY 1979

Prepared for

U.S. DEPARTMENT OF TRANSPORTATION
United States Coast Guard
Office of Research and Development
Washington, D.C. 20590

NOTICE

This document is disseminated under the sponsorship of the Department of Transportation in the interest of information exchange. The United States Government assumes no liability for its contents or use thereof.

The contents of this report do not necessarily reflect the official view or policy of the Coast Guard; and they do not constitute a standard, specification, or regulation.

This report, or portions thereof may not be used for advertising or sales promotion purposes. Citation of trade names and manufacturers does not constitute endorsement or approval of such products.

(18) USCG (19) D-16-79

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
CG-D-16-79		

(6) **4. Title and Subtitle**
Evaluation and Classification of the Electrical Hazard of Chemical Vapors During Water Transportation Using Pattern Recognition Techniques

5. Report Date
JAN 1979

6. Performing Organization Code

8. Performing Organization Report No.

(10) **7. Author(s)** James L. Fasching, Earl W. Stromberg and Clifford P. Weisel

9. Performing Organization Name and Address
Department of Chemistry
University of Rhode Island
Kingston, Rhode Island 02881

(11) Jan 79

10. Work Unit No. (TRAIS)

11. Contract or Grant No.
DOT-CG-44160-A

13. Type of Report and Period Covered
(9) Final Report

12. Sponsoring Agency Name and Address
U. S. Coast Guard
Office of Research and Development
Washington, D. C. 20590

410 237

14. Sponsoring Agency Code

15. Supplementary Notes
The U. S. Coast Guard Office of Research and Development technical representative for the work performed herein was Dr. Michael Parnarouskis.

16. Abstract

Pattern recognition is a statistical technique that allows one to find or predict a property of chemicals that is not directly measurable, but is known to depend upon certain features or properties of the chemicals via some totally unknown relationship. This technique has been applied to a multitude of scientific problems. The same technique was used to classify a chemical according to its relative hazard in bulk water-transportation based on chemical structure and macro-scale properties such as density, vapor pressure, structure-fragments, solubilities, etc.

Using the Linear-Learning Machine, the overall prediction of the 47 compounds in training set was 68% correct. The predicted classifications of the 240 compounds in the test set are approximately 68% correct. There are many difficulties associated with properly classifying compounds on the basis of variable derived from structural fragments that must be solved before great reliance can be placed on the results of a Linear-Learning Machine classification.

(15) DOT-CG-44160-A (12) 90 p.

17. Key Words	18. Distribution Statement
Electrical Hazard, Chemical Vapors, Water Transport, Pattern Recognition	Document is Available to the public through the National Technical Information Service, Springfield, Va. 22161.

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	84	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

410 237

A

METRIC CONVERSION FACTORS

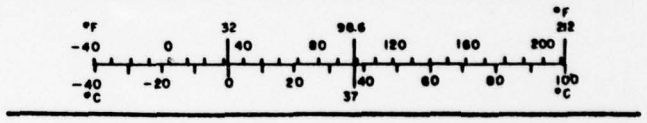
Approximate Conversions to Metric Measures

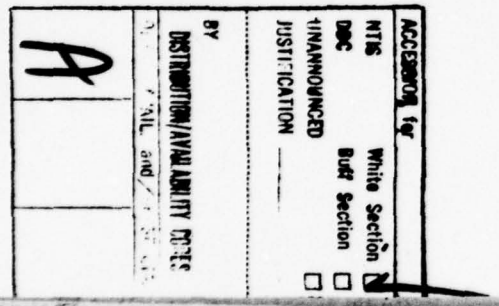
Symbol	When You Know	Multiply by	To Find	Symbol
	LENGTH			
in	inches	*2.5	centimeters	cm
ft	feet	30	centimeters	cm
yd	yards	0.9	meters	m
mi	miles	1.6	kilometers	km
	AREA			
in^2	square inches	6.5	square centimeters	cm^2
ft^2	square feet	0.09	square meters	m^2
yd^2	square yards	0.8	square meters	m^2
mi^2	square miles	2.6	square kilometers	km^2
	acres	0.4	hectares	ha
	MASS (weight)			
oz	ounces	28	grams	g
lb	pounds	0.45	kilograms	kg
	short tons (2000 lb)	0.9	tonnes	t
	VOLUME			
tsp	teaspoons	5	milliliters	ml
Tbsp	tablespoons	15	milliliters	ml
fl oz	fluid ounces	30	milliliters	ml
c	cups	0.24	liters	l
pt	pints	0.47	liters	l
qt	quarts	0.95	liters	l
gal	gallons	3.8	liters	l
ft^3	cubic feet	0.03	cubic meters	m^3
yd^3	cubic yards	0.76	cubic meters	m^3
	TEMPERATURE (exact)			
°F	Fahrenheit temperature	5/9 (after subtracting 32)	Celsius temperature	°C

*1 in = 2.54 (exactly). For other exact conversions and more detailed tables, see NBS Misc. Publ. 286, Units of Weights and Measures, Price $2.25, SD Catalog No. C13.10:286.

Approximate Conversions from Metric Measures

Symbol	When You Know	Multiply by	To Find	Symbol
	LENGTH			
mm	millimeters	0.04	inches	in
cm	centimeters	0.4	inches	in
m	meters	3.3	feet	ft
m	meters	1.1	yards	yd
km	kilometers	0.6	miles	mi
	AREA			
cm^2	square centimeters	0.16	square inches	in^2
m^2	square meters	1.2	square yards	yd^2
km^2	square kilometers	0.4	square miles	mi^2
ha	hectares (10,000 m^2)	2.5	acres	
	MASS (weight)			
g	grams	0.035	ounces	oz
kg	kilograms	2.2	pounds	lb
t	tonnes (1000 kg)	1.1	short tons	
	VOLUME			
ml	milliliters	0.03	fluid ounces	fl oz
l	liters	2.1	pints	pt
l	liters	1.06	quarts	qt
l	liters	0.26	gallons	gal
m^3	cubic meters	35	cubic feet	ft^3
m^3	cubic meters	1.3	cubic yards	yd^3
	TEMPERATURE (exact)			
°C	Celsius temperature	9/5 (then add 32)	Fahrenheit temperature	°F





Department of Chemistry
Pastore Laboratory
University of Rhode Island
Kingston, Rhode Island 02881

Evaluation and Classification of the Electrical Hazard of Chemical Vapors During Water Transportation Using Pattern Recognition Techniques

Final Report of Contract DOT-CG-44160-A Submitted to the United States Coast Guard, Department of Transportation

Revised Version

Submitted by:

James L. Fasching
Earl W. Stromberg
Clifford P. Weisel

TABLE OF CONTENTS

1. ABSTRACT

Pattern recognition is a statistical technique that allows one to find or predict a property of chemicals that is not directly measurable, but is known to depend upon certain features or properties of the chemicals via some totally unknown relationship. This technique has been applied to a multitude of scientific problems. The same technique was used to classify a chemical according to its relative hazard in bulk water-transportation based on chemical structure and macro-scale properties such as density, vapor pressure, structure-fragments, solubilities, etc.

Using the Linear-Learning Machine, the overall prediction of the 47 compounds in training set was 68% correct. The predicted classifications of the 240 compounds in the test set are approximately 68% correct. There are many difficulties associated with properly classifying compounds on the basis of variables derived from structural fragments that must be solved before great reliance can be placed on the results of a Linear-Learning Machine classification.

2. INTRODUCTION

The U. S. chemical industry transports the bulk of its chemical feedstocks and products by water. This movement of large amounts of chemicals by tankers, barges, etc., constitutes a definite fire, health and poison hazard as well as possible physiological irritants and water pollutants if spills occur (1). The U. S. Coast Guard has primary responsibility for the safety of shipping, waterways and citizens of this nation. Some methods of assessing the hazards of these chemicals during bulk transportation by water have been developed by various organizations (1-5) for the U. S. Coast Guard. The National Research Council's Committee on Hazardous Materials (Division of Chemistry and Chemical Technology) (1) has issued a Tentative Guide for the Evaluation of the Hazard of Bulk Water Transportation of Industrial Chemicals which outlines a system of evaluation. It also tentatively rates 337 common industrial chemicals.

The Fire Hazard aspects of this overall problem has deeply concerned the Coast Guard because of the potential loss of people, ship and damage to the environment. Underwriters' Laboratories, Inc. have tested 53 chemicals determining the flame propagation effects and pressure piling developed by various gas and/or vapor-air mixtures of these chemicals. They used the Westerberg Explosion Test Vessel which measures flash points, ignition temperatures and flammability limits of these chemicals. Using this data, the Electrical Hazards Panel of the Committee on Hazardous Materials (the Division of Chemistry and Chemical Technology, National Research Council) has tentatively classified 370 chemicals as to their relative fire hazard with respect to explosion-proof electrical equipment. These tentative classifications are based on the experimental data from the subset of 53 chemicals that have been tested by Underwriters' Laboratories, Inc. The current classifications for a large number (>200) of chemical compounds are essentially an educated guess by a panel of experts.

3. RESEARCH OBJECTIVES

The overall objective of the research conducted under this contract was to utilize pattern recognition techniques to develop a computer program that would quickly, cheaply and effectively evaluate a chemical compound as to its fire-hazard classification for bulk water transport. The information to perform this classification would be obtained from the chemical structure and other simple chemical-physical properties of the compound.

Chemometrics, a growing discipline in chemistry, can be defined as the study of new mathematical and statistical approaches to solving chemical problems. Pattern recognition (9), a subset of these mathematical methods, has recently been applied to many chemical problems. Recent reviews (8,9) reference much of the literature that demonstrates the unique adaptations of pattern recognition techniques to solve problems in chemistry. The same techniques have been applied herein to the problem of classifying a chemical according to its relative fire hazard during bulk water transportation.

4. SOFTWARE DEVELOPMENT

Almost 50% of our effort was spent on developing computer codes for use on the fire hazard classification problem. A generalized factor analysis program, 3-dimensional plotting and hierarchial clustering routines were written for use on IBM-370/155. These programs proved to be of marginal use in solving the problem. A program named ARTHUR (12) written by Duewer, Harper and Kowaliski from the University of Washington and Fasching, Weisel and Stromberg from the University of Rhode Island was used to obtain the data presented in this report. Appendix A and B explain in detail the terms, behavior and capabilities of these programs.

5. EVALUATION OF VARIOUS VARIABLES

The objective of this examination was to find the fewest number of variables that gave complete separation of the compounds in the training set according to their category. The variables were chosen because they give distinguishing characteristics about each compound's reactivity, combustibility, vapor pressure and/or some other property which might contribute to its fire hazard. All the variables being used are either chemical or physical values that are measurable quantities as either distinct numbers or categorized from experimental data. Using measured quantities avoids any biases that would result from data that is subject to change because of extrapolation by the scientist. Variables for which no experimental data was found were estimated by using a range of various values based on other similar compounds and general trends for that variable.

The variable order (Table I) was determined using histograms (Fig.1); two features of the linear learning machines in ARTHUR described below; step-wise discriminate function routines in the BMDP package (12); and the Fisher-Weight, Variance Weight and Property Weight step-wise discriminate features of the SELECT routine in ARTHUR (Appendix II). Step-wise discriminate methods assume that one can separate out variables on the basis of decreasing importance with respect to variance.

The histograms allow us to determine which variable has the least amount of association, i.e. is the most independent. The two features listed below describe how the linear learning machine was used to determine the efficiency of a set of variables in predicting the results. They are: the smaller the number of passes made the quicker the results converged and if 100% separation was not obtained which compound was incorrectly classified. The second feature is useful since an examination of the type of compounds incorrectly classified in the training set shows whether too little or too much emphasis is being included about particular classes

of compounds or functional groups. The result of our variable selection on the training set is listed in Table V and Table VI. Table V shows that complete separation of the compounds into the appropriate categories is obtained when all compounds are included. Table VI is a compilation of the results of a JACKKNIFE study. The JACKKNIFE procedure uses one of known compounds as a test case and compares the predicted value with the experimentally determined value. The complete training set is treated in this fashion. When all of the compounds are used 100% separation into the correct categories is obtained. In JACKKNIFE this was not the case. For example when 2-nitropropane is used as a test compound, therefore, left out of training set and not included in the determination of the eigenvectors, it is incorrectly associated with the D group. Since 2-nitropropane is the only compound in the training set containing a NO_2 group, it is probably that unique characteristic that aids in the compound's classification when included in the training set (Table V). 2-nitropropane, therefore seems to be able to contribute useful information upon which decisions about other NO_2 containing compounds can then be made. Step-wise discriminate analysis (described in Appendix II) determines the order of importance of variables according to how well a variable retains the compounds in the appropriate category.

The reasons these routines yielded the ordering found in Table I is not very obvious as it depends on complex interactions among the variables. The ordering is based on empirical results. A simple case of this is that total chlorine and the number of chlorines attached to carbon seems to have very different importance as they are listed as variable numbers 17 and 39 respectively, although one would guess that they should have very similar importance. Actually in the data set chosen they are exactly equivalent, i.e. all chlorines present are attached to carbons. The step-wise discriminant program in the BMDP package and those in the ARTHUR package when presented with two variables that are equivalent or one that is a

linear combination of another or sets of others, recognize this fact and essentially eliminates one of the variables by making its contribution to the classifying function negligable, thereby eliminating duplicate information. In addition to eliminating duplicate information differences in the unit size among variables was standardized by autoscaling to unit variance and zero mean. This feature normalizes differences between variables due to units, thus making it possible to compare values such as temperature, solubility and number of functional groups present. It also makes it inconsequential what units one chooses for measuring any variable, such as ignition temperatures, whether it be degrees kelvin, centigrade or farenheit, provided the same unit is used for all compounds.

Table IV lists the 13 variables that were found to give the optimal results. An examination of Table IV shows that a combination of variables that contain information distributed among all compounds, such as AIT, molecular weight and solubilities must be coupled with information about specific functional groups such as epoxy, nitro and NH groups. We again wish to mention that if any large family of compounds containing one type of functional group is eventually going to be classified, it is important to have a few examples of the family in the training set to see if that functional group contributes to its classification or if its effects can be accounted for by other variables present. The reason for the classifying function containing both variables which contain values for all compounds and variables that only a few compounds that have a value other than zero is the following: the first type of variable sets up a basis where very general trends are found, such as high molecular weight compounds are less hazardous than lighter ones. The second type, functional groups, are needed since they can activate or deactivate the reactivity of a compound greatly, thus shifting its classification.

6 CLASSIFICATION PREDICTION

The variables list generated, as described in the previous section, was utilized to predict the categories of unknowns. First, we tried to separate the classified compounds into their appropriate categories and minimize the number of variables needed for the optimal results. This was then checked to estimate the accuracy of our prediction and then was applied to unknowns. A discussion of the procedures used to accomplish these steps follows.

All of the compounds that have been experimentally classified into NAS groups B, C, and D at temperatures of less than 25°C were used in this research. (These are listed in Table V) The two compounds classified into the A group, acetylene and carbon disulfide are included into category B. This produces a group of compound that include hazardous classification B or above. The reason for this is that only two compounds do not provide an adequate mathematical basis to establish a legitimate pattern to distinguish which variables are instrumental in classifying compounds or form a base for predicting unknown compounds. Leaving the two compounds out completely does not markly effect the B groups.

As mentioned above, only compounds classified experimentally below 25°C are being utilized. Those whose data were obtained above 25°C are not being included since their classification at the normalized temperature may not be the same and therefore would contribute inaccurate information to the list of variable values for those compounds. The methyl acetylene-propadiene (MAPP) gas mixture and gasoline mixtures are also not being used, since unique chemical and physical properties cannot be defined for these substances. It should be noted that the classifying method we are attempting to develop cannot be used for any mixtures or heterogeneous substances.

The BMDP program used requires that a priority of weighting be set between the categories. This value should reflect both the suspected "cost" of the misclassifying of a compound that actually belongs to the group B

into the group C, B into D, etc., and the probability of a randomly selected compound being either a B, C, or D. A number of different priorities were used with the ratio of B to C to D of 1 to 2 to 1 most prevalent. This ratio was calculated by estimating the cost of misclassification of a B into a C and a C into a D category of five times (John M. Cece, private communication, 1976) since it is thought in the case of an accident the inadequate safeguards would cause a higher loss of both life and material than the expense of using a higher safeguard system for a less dangerous compound. The probability of random selection was calculated by dividing the total number of compounds in each category by the total number of compounds listed in the report entitled "Matrix of Electrical and Fire Hazard Properties and Classification of Chemicals" by the Committee of Hazardous Materials. It was assumed that this report contained a large, randomly selected samples of chemicals that will be shipped and that a reasonable number of assigned categories are correct (.75%).

Table V is a list of the 47 chemical compounds that were used as a training set for the various pattern recognition techniques. Only compounds that were not experimentally tested at elevated temperatures were included. A training set is a group of compounds that the programs assume to be correct and is used to train the program to predict classification of test samples. The two principal learning machines that were finally used to solve the fire hazard classification problem were PLANE and MULTI. The programs can train themselves to be 100% correct with respect to the training set. (Appendix II)

Table VI is a list of the same training set shown in Table V, but each compound was considered to be a test set respectively. Forty-seven computer runs were subsequently performed using the JACKKNIFE procedure described previously with the 46 chemicals being the training set. The results of these experiments are summarized in Table VII. The programs

PLANE and MULTI correctly predict the training set ~70% of the time when each compound in the training set is considered as a test compound.

A summary of the compounds that were misclassified is given, by category, in Table VII. Category D was predicted at a reliability between 74% and 79%, category C at 67% and category B at 43%. A possible explanation for the misclassifications of the experimentally determined compounds are derived from their spark gap values. Boundaries of spark gap values of 0.010" and 0.030" are the apparent divisions between the B and C categories and C and D categories respectively. There are a few compounds that are classified in a category other than would be suggested by these cutoffs due to their anomalous high pressure piling values (>250 psig). We have predicted a number of these compounds to be different from the assigned category. They are ethylene diamine, vinyl chloride, cyclopropane, 1,3 butadiene and propylene. An additional number of compounds within 0.005" of the spark gap boundaries are also misclassified (Table III). Two compounds with whose category we have agreed have spark gap values on the "wrong" side of the boundary. They are isoprene, classified as a C with a spark gap values of 0.037", and acrolein, classified as a B with a spark gap value of 0.018". It seems probable, therefore, that some of the problems we have had could be due to the boundaries between categories not being clear cut and pressure piling values only being considered when they are very large.

Three of the misclassified compounds in the training set (hydrogen, hydrogen sulfide and carbon disulfide) are the only inorganic compounds present. Since organic and inorganic compounds do not react the same way chemically, it is possible that these were placed in an incorrect category since we based their prediction on information derived from organic compounds.

The size of the training set and the uniqueness of certain characteristics present in the chemical being tested may result in incorrect classification. If a characteristic is not present in any compounds remaining in the training set yet contributes greatly to the fire hazard of the chemical tested, an incorrect result would occur. The classification would therefore be based on other features of the compound because there is no information about that unique feature.

Table IX lists the test compounds that have been classified into a higher category than in current use (5). Alcohols, long chain alkanes, alkenes and benzene substituted compounds comprise a large portion of this table. A possible explanation for our higher classification of these four types of compounds may be the result of insufficient data because of similar compounds are not found in the training set, and/or training set compounds in these groups are being classified in the higher category.

7. CONCLUSIONS

The results for the training set of 47 compounds is summarized in Table VII. The two linear learning machine called PLANE and MULTI were used to classify the compounds. The predicted classifications in Table VII consider each chemical to be a test sample and the routines are trained on the other 46 chemicals. This set of analysis gives an overall estimate of 68% correct classification. Table VIII lists the results of PLANE and MULTI when the 47 compounds are used as the training set and the 240 compounds are used as a test set. The predicted classifications are approximately 68% correct. Almost all long chain and short chain alcohols are classified one category higher by both routines. Table IX lists the compounds that have been classified upwards one category.

The final results of this research are encouraging and they have definitely indicated the direction of future research.

8 RECOMMENDATIONS

The difficulties of properly classifying compounds on the basis of variables derived from structural fragments has been well documented in this report and current literature.

It is our opinion that these results could be improved if the following four areas were given further consideration:

1. Use pattern recognition techniques based on a property descriptor instead of Classes (A, B, C, D). We can predict measured experimental variables such as AIT, pressure piling etc. much better than abstract classifications that are based on these variables. We would also have the ability to accurately check our results using this approach and scientists would make the final evaluation.
2. Explore the use of general molecular descriptors, thermodynamic properties and electron density parameters as variables in the pattern recognition techniques. Such variables would be obtained from molecular orbital programs, CHETAH etc. and used in this work to improve the prediction capability.
3. Devise better feature extraction procedures and transformations to eliminate the random or noise components. A major problem in pattern recognition concerns the mathematical ability to separate noise from real and useful information in a variable. Better techniques for this problem must be found.
4. Design a better learning machine program. The learning machine has many faults but one great advantage. It is extremely simple to use after it is trained. In fact, it can be done on a simple calculator. Its major problem (accuracy) has been well documented in the report. If program modifications could be developed to solve these problems, the pattern recognition field would make a major leap forward.

TABLE I

Variable Ordering

1. Auto ignition temperature
2. Total number of hydrogens
3. Epoxy groups
4. NO_2 groups
5. Molecular weight
6. Solubility in ether
7. CH_3 groups
8. NH_2 groups
9. Total number of carbons
10. Carbon-carbon single bonds
11. Carbon-carbon triple bonds
12. Ether linkages
13. Total number of sulfurs
14. NH groups
15. Solubility in alcohol
16. Carbon-chlorine bonds
17. NH_3 groups
18. Carbons without hydrogens
19. Carbon-nitrogen triple bonds
20. CH groups
21. N-C=N groups
22. Ethyl groups
23. COH groups
24. Total number of nitrogens
25. Hydrogens alpha to C=O
26. Ester linkages
27. Nitrogens without hydrogens
28. Solubility in water
29. Hydrogens alpha to C=C
30. Flash point
31. Total number of oxygens
32. Boiling point
33. HC=O groups
34. Melting point
35. CH_2 groups
36. COOH groups
37. C=O groups
38. Carbon-carbon double bonds
39. Total number of chlorines

TABLE II

CLASSIFICATION FUNCTIONS[a]

VARIABLE #	GROUP D	GROUP C	GROUP B
1 X(1)	7.13600	4.97641	2.76057
2 X(2)	-0.36106	-0.43583	-0.40036
4 X(4)	-0.06152	0.12287	0.10499
5 X(5)	20.51698	8.62925	10.51797
6 X(6)	12.02098	20.64725	18.11748
7 X(7)	6.67528	6.92967	3.76196
8 X(8)	4.35670	-6.77126	-6.94182
10 X(10)	-17.96806	-31.78358	-26.15622
11 X(11)	-37.05293	-22.64537	-7.63201
14 X(14)	53.06761	115.01823	112.80058
15 X(15)	-31.01567	-25.29327	-15.60228
16 X(16)	-26.66524	11.50037	29.91949
17 X(17)	-21.73500	53.08704	61.99918
19 X(19)	-3.70266	5.61196	12.42352
20 X(20)	-47.82832	56.20058	58.74736
21 X(21)	-48.94760	33.24651	30.70747
23 X(23)	-53.36461	-145.65991	-10[illegible].30605
24 X(24)	-86.22984	-109.46097	-89.71614
26 X(26)	-146.34658	-104.83170	-70.20551
27 X(27)	-43.60481	-113.88313	-73.73346
30 X(30)	-25.06001	-14.32377	-3.85582
31 X(31)	-66.89027	-13.61657	6.76443
32 X(32)	-13.45770	-13.57369	2.25790
33 X(33)	-182.14958	-114.14978	-40.13753
34 X(34)	-54.96423	-50.60464	-66.21277
35 X(35)	-85.81082	23.76089	55.32253
36 X(36)	0.45324	0.41791	0.42203
37 X(37)	-4.26777	-9.11450	-6.22199
38 X(38)	-3.09302	-4.28917	-2.83177
39 X(39)	-33.49681	-17.37549	6.43740
CONSTANT	-224.74899	-199.83459	-188.00497

[a]Each numerical value in the table is the coefficient of a linear polynomial of the 39 variables plus the constant. For example, $y = a_0+a_1X_1+a_2X_2+a_3X_3+\ldots$ The calculated values of y for each compound can be used to determine the appropriate classification.

TABLE II

CLASSIFICATION FUNCTIONS[a]

VARIABLE #	GROUP D	GROUP C	GROUP B
1 X(1)	7.13600	4.97641	2.76057
2 X(2)	-0.36106	-0.43583	-0.40036
4 X(4)	-0.06152	0.12287	0.10499
5 X(5)	20.51698	8.62925	10.51797
6 X(6)	12.02098	20.64725	18.11748
7 X(7)	6.67528	6.92967	3.76196
8 X(8)	4.35670	-6.77126	-6.94182
10 X(10)	-17.96806	-31.78358	-26.15622
11 X(11)	-37.05293	-22.64537	-7.63201
14 X(14)	53.06761	115.01823	112.80058
15 X(15)	-31.01567	-25.29327	-15.60228
16 X(16)	-26.66524	11.50037	29.91949
17 X(17)	-21.73500	53.08704	61.99918
19 X(19)	-3.70266	5.61196	12.42352
20 X(20)	-47.82832	56.20058	58.74736
21 X(21)	-48.94760	33.24651	30.70747
23 X(23)	-53.36461	-145.65991	-107.30605
24 X(24)	-86.22984	-109.46097	-89.71614
26 X(26)	-146.34658	-104.83170	-70.20551
27 X(27)	-43.60481	-113.88313	-73.73346
30 X(30)	-25.06001	-14.32377	-3.85582
31 X(31)	-66.89027	-13.61657	6.76443
32 X(32)	-13.45770	-13.57369	2.25790
33 X(33)	-182.14958	-114.14978	-40.13753
34 X(34)	-54.96423	-50.60464	-66.21277
35 X(35)	-85.81082	23.76089	55.32253
36 X(36)	0.45324	0.41791	0.42203
37 X(37)	-4.26777	-9.11450	-6.22199
38 X(38)	-3.09302	-4.28917	-2.83177
39 X(39)	-33.49681	-17.37549	6.43740
CONSTANT	-224.74899	-199.83459	-188.00497

[a]Each numerical value in the table is the coefficient of a linear polynomial of the 39 variables plus the constant. For example, $y = a_0 + a_1X_1 + a_2X_2 + a_3X_3 + \ldots$ The calculated values of y for each compound can be used to determine the appropriate classification.

TABLE III

Compound	NAS Classification	Spark[a] Gaps (inches)	Pressure Piling (lbs/sq. in.)	Borderline[b] Compounds
Methane	D	.044	77	
Ethylene Diamine	D	.029	82	*
Ethylamine	D	.039	65	
Styrene	D	.037	133	
Vinyl Acetate	D	.041	128	
Vinyl Chloride	D	.029	165	*
Allyl Alcohol	C	.026	120	*
Epichlorohydrin	C	.022	149	
Hydrogen Sulfide	C	.026	60	*
2-Nitropropane	C	.021	130	
Triethylamine	C	.021	125	
Cyclopropane	C	.034	147	*
Methyl Acetylene	C	.025	185	*
Ethylene	C	.027	180	*
1,3 Butadiene	B	.031	260	*
Carbon Disulfide	B	.002	205	
Propylene Oxide	B	.021	280	*
Hydrogen	B	.003	845	

[a]Spark gap tentative standards are less than 0.010" for A, B between 0.010" and 0.030" for C, and greater than 0.030" for D.

[b]Compounds with sparks $\pm$0.005 of the tentative standards are marked in this column.

TABLE IV

VARIABLES USED

(1) Molecular weight

(2) Solubility in ether

(3) Solubility in alcohol

(4) CH_3 group

(5) Carbon-carbon single bonds

(6) NH groups

(7) NH_2 groups

(8) NO_2 groups

(9) Ester linkages

(10) Total number of carbons

(11) Total number of hydrogens

(12) Auto ignition temperatures

(13) Epoxy groups

TABLE V

TRAINING SET

COMPOUNDS	NAS	MULTI	PLANE**
Allyl Alcohol	C	C	C
Acrolein	B	B	B
sec-Butyl Alcohol	D	D	D
n-Butyl Aldehyde	C	C	C
Crotonaldehyde	C	C	C
Diethylamine	C	C	C
Diisobutylene	D	D	D
Epichlorohydrin	C	C	C
Ethyl Acrylate	D	D	D
Ethylene Diamine	D	D	D
Ethyleneimine	C	C	C
Hydrogen Sulfide	C	C	C
Isopropyl Ether	D	D	D
Mesityl Oxide	D	D	D
Morpholine	C	C	C
2-Nitropropane	C	C	C
Pyridine	D	D	D
Tetrahydrofuran	C	C	C
Methane	D	D	D
Methyl Formal	C	C	C
Dimethyl Ether	C	C	C
Di-n-Propyl Ether	C	C	C
Ethylamine	D	D	D
Triethylamine	C	C	C
Cyclopropane	C	C	C
Methyl Acetylene	C	C	C
Propane	D	D	D
Acetaldehyde	C	C	C
Acrylonitrile	D	D	D
Ammonia	D	D	D

criminant program in the BMDP package and those in the ARTHUR package when presented with two variables that are equivalent or one that is a

Table V cont'd.

COMPOUNDS	NAS	MULTI	PLANE**
1,3 Butadiene	B	B	B
Carbon Disulfide	B*	B	B
Ethylene Dichloride	D	D	D
Ethylene Oxide	B	B	B
Isoprene	C	C	C
Propylene	D	D	D
Propylene Oxide	B	B	B
Styrene	D	D	D
Unsymmetric Dimethyl-Hydrazine (UDMH)	C	C	C
Vinyl Acetate	D	D	D
Vinyl Chloride	D	D	D
Para-Xylene	D	D	D
Hydrogen	B	B	B
Diethyl Ether	C	C	C
Ethylene	C	C	C
Butane	D	D	D
Acetylene	B*	B	B

*Compounds in A category with the B category.

**PLANE decides between two categories.

can activate or deactivate the reactivity of a compound greatly, thus influencing its classification.

TABLE VI

TRAINING SET CLASSIFICATIONS

COMPOUNDS	NAS	MULTI	PLANE**
Allyl Alcohol	C	C	D_C
Acrolein	B	B	B_C
sec-Butyl Alcohol	D	D	D_C
n-Butyl Aldehyde	C	C	C_D
Crotonaldehyde	C	C	C_D
Diethylamine	C	C	C_D
Diisobutylene	D	D	D_C
Epichlorohydrin	C	B	B_D
Ethyl Acrylate	D	D	D_C
Ethylene Diamine	D	B	B_C
Ethyleneimine	C	C	C_B
Hydrogen Sulfide	C	B	B_C
Isopropyl Ether	D	D	D_C
Mesityl Oxide	D	D	D_C
Morpholine	C	C	C_D
2-Nitropropane	C	D	D_C
Pyridine	D	D	D_C
Tetrahydrofuran	C	C	C_D
Methane	D	C	C_D
Methyl Formal	C	C	C_B
Dimethyl Ether	C	C	C_D
Di-n-Propyl Ether	C	C	C_D
Ethylamine	D	C	C_D
Triethylamine	C	B	D_C

tween the categories. This value should reflect both the suspected "cost" of the misclassifying of a compound that actually belongs to the group B

TABLE VI cont'd.

COMPOUNDS	NAS	MULTI	PLANE**
Cyclopropane	C	D	C_D
Methyl Acetylene	C	D	D_C
Propane	D	D	D_C
Acetaldehyde	C	C	C_D
Acrylonitrile	D	D	D_C
Ammonia	D	D	D_C
1,3 Butadiene	B	C	Unclassified***
Carbon Disulfide	B*	C	C_B
Ethylene Dichloride	D	D	D_C
Ethylene Oxide	B	B	B_C
Isoprene	C	C	C_B
Propylene	D	D	D_C
Propylene Oxide	B	C	C_D
Styrene	D	B	D_B
Unsymmetric Dimethyl Hydrazine (UDMH)	C	C	C_D
Vinyl Acetate	D	D	C_B
Vinyl Chloride	D	D	C_D
Para-Xylene	D	D	D_C
Hydrogen	C	C	C_D
Diethyl Ether	C	C	C_D
Ethylene	C	B	B_C
Butane	D	D	D_C
Acetylene	B*	B	B_C

*Compounds in A category grouped with the B category.

**PLANE is a two category classifier, with the subscript being the category choice between the two categories originally not selected by plane.

***PLANE, which examines only two groups at a time, did not give a unique answer for all three pairs.

described previously with the 46 chemicals being the training set. The results of these experiments are summarized in Table VII. The programs

TABLE VII
INCORRECTLY CLASSIFIED

MULTI[a]

Category	D	C	B
	Methane (C)	Epichlorohydrin (B)	1,3 Butadiene (C)
	Ethylene Diamine (B)	Hydrogen Sulfide (B)	Carbon Disulfide (C)
	Ethylamine (C)	2-Nitropropane (D)	Propylene Oxide (C)
	Styrene (B)	Triethylamine (B)	Hydrogen (C)
		Cyclopropane (D)	
		Methyl Acetylene (D)	
		Ethylene (B)	
	15/19 = 79%	14/21 = 67%	3/7 = 43%
	Total 32/47 = 68%		

PLANE[a]

Category	D	C	B
	Methane (C)	Allyl Alcohol (D)	1,3 Butadiene (C)
	Ethylene Diamine (B)	Epichlorohydrin (B)	Carbon Disulfide (C)
	Ethylamine (C)	Hydrogen Sulfide (B)	Propylene Oxide (C)
	Vinyl Acetate (C)	2-Nitropropane (D)	Hydrogen (C)
	Vinyl Chloride (C)	Triethylamine (D)	
		Methyl Acetyene (D)	
		Ethylene (B)	
	14/19 = 74%	14/21 = 67%	3/7 = 43%
	Total 31/47 = 66%		

[a]Appendix II

TABLE VIII

Test Compounds taken from "Matrix of Electrical and Fire Hazard Properties and Classification of Chemicals"

Number	Compound Name	Classification		
		NAS	MULTI[a]	PLANE[a]
1	Formic Acid	D	D	D_C
2	Acetic Acid	D	D	D_C
3	Propionic Acid	D	D	D_C
4	n-Butyric Acid	D	D	D_C
5	Acrylic Acid*	D	C	D-C
6	Acetic Anhydride	D	D	D_C
7	Propionic Anhydride	D	D	Unclassified**
8	Phthalic Anhydride	D	D	Unclassified**
9	Methyl Alcohol	D	C	D_C
10	Ethyl Alcohol	D	C	C_D
11	n-Propyl Alcohol	D	D	D_C
12	iso-Propyl Alcohol	D	C	C_D
13	n-Butyl Alcohol	D	C	C_D
14	sec-Butyl Alcohol	D	Training	
15	iso-Butyl Alcohol	D	D	D_C
16	tert-Butyl Alcohol	D	D	D_C
17	n-Amyl Alcohol	D	C	C_D
18	iso-Amyl Alcohol	D	C	C_D
19	Hexanol	D	C	C_D
20	Methylamyl Alcohol*	D	C	C_D
21	Methyl Isobutyl Alcohol*	D	C	D-C
22	Ethyl Butanol*	D	C	D-C
23	Cycohexanol	D	C	C_B

TABLE VIII cont'd.

Number	Compound Name	TEST COMPOUNDS Classification NAS	MULTI	PLANE
24	n-Octyl Alcohol*	D	C	D_C
25	iso-Octyl Alcohol*	D	C	D_C
26	2-Ethyl Hexanol*	D	C	D_C
27	Nonyl Alcohol*	D	C	D_C
28	Diisobutyl Carbonal*	D	C	D_C
29	n-Decyl Alcohol	D	C	D_C
30	iso-Decyl Alcohol*	D	C	D_C
31	Undecanol*	D	C	D_C
32	Dodecanol	D	C	D_C
33	Tridecanol*	D	C	D_C
34	Tetradecanol*	D	C	D_C
35	Pentadecanol*	D	C	D_C
36	Allyl Alcohol	C	Training	
37	Diacetone Alcohol	D	D	D_C
38	Formaldehyde Solution	-	C(pure)	C_B(pure)
39	Acetaldehyde	C	Training	
40	Propionaldehyde	C	C	C_D
41	n-Butyraldehyde	C	Training	
42	iso-Butyraldehyde*	C	C	C_D
43	Valeraldehyde*	C	C	C_{D-B}
44	3-Methyl Butyraldehyde*	C	C	C_D
45	iso-Pentyl Aldehyde*	C	C-B	C_D
46	2-Ethylhexaldehyde	C	B	C_D
47	iso-Octyl Aldehyde*	C	B	C_D
48	n-Decaldehyde*	C	C	C_B
49	iso-Decaldehyde*	C	C	C_D
50	Acolein	B	Training	
51	Crotonaldehyde	C	Training	

TABLE VIII cont'd. TEST COMPOUNDS

Number	Compound Name	Classification NAS	MULTI	PLANE
52	2-Ethyl-3-Propyl Acrolein*	C	C	C_D
53	Glyoxal*	C	B	B
54	Glutaraldehyde*	C	C	C_D
55	Furfural	C	C	C_D
56	Methane	D	Training	
57	Ethane	D	D	D_C
58	Propane	D	Training	
59	Butane	D	Training	
60	n-Pentane	D	C	C_D
61	iso-Pentane	D	D	D_C
62	n-Hexane	D	C	C_D
63	iso-Hexane	D	C	C_D
64	n-Heptane	D	C	C_D
65	Octane	D	C	C_D
66	Nonane	D	C	C_D
67	Cyclopropane	C	Training	
68	Cyclohexane	D	C	C
69	Monoethanolamine*	D	D	D_B
70	Diethanolamine	D	C	C_D
71	Triethanolamine*	D	D	D_C
72	Monaisopropanolamine*	D	D	D_C
73	Diisopropanolamine*	D	C	D_C
74	n-Aminoethyl Ethanolamine	D	C	C_D
75	Ethylamine	D	Training	
76	iso-Propylamine	D	D	D_C
77	Dimethylamine	C	C	C_D

TABLE VIII cont'd. TEST COMPOUNDS

Number	Compound Name	NAS	Classifications MULTI	PLANE
78	Diethylamine	C	Training	
79	Di-n-propylamine*	C	C	C_D
80	Diisopropylamine*	C	C	D_C
81	Triethylamine	C	Training	
82	Ethylene Diamine	D	Training	
83	Hexamethylene Diamine Solutions*	D	D-B	D_C
84	Diethylenetriamine	D	C	C_D
85	Triethylene Tetramine*	D	C	D_C
86	Tetraethylene Pentamine*	D	C	D_C
87	Ethylenimine	C	Training	
88	Hexamethylenimine*	C	C	C_D
89	Aniline	D	D	D_C
90	Pyridine	D	Training	
91	2-Methyl-5-Ethyl Pyrdine*	D	D	D_C
92	Benzene	D	D	D_C
93	Toluene	D	D	D_C
94	Ethyl Benzene	D	D	D_C
95	Cumene	D	D	D_C
96	Decyl Benzene*	D	C	D_C
97	Undecyl Benzene*	D	C	D_C
98	Dodecyl Benzene*	D	C	D_C
99	Tridecyl Benzene*	D	C	D_C
100	Tetradecyl Benzene*	D	C	D_C
101	o-Xylene	D	D	D_C
102	m-Xylene	D	D	D_C
103	p-Xylene	D	Training	

TABLE VIII cont'd. TEST COMPOUNDS

Number	Compound Name	Classifications NAS	MULTI	PLANE
104	Xylene (mixture)	D	-	-
105	p-Cymene	D	D	D_C
106	Diethylbenzene	D	D	D_C
107	Triethyl Benzene*	D	D	D_C
108	Styrene	D	Training	
109	Vinyl Toulene*	D	D	D_C
110	Naphthalene	D	D	D_B
111	Tetrahydronaphthalene	D	D	D_C
112	Mixture	D	-	-
113	Methyl Acetate	D	D	D_C
114	Ethyl Acetate	D	D	D_C
115	n-Propyl Acetate	D	D	D_C
116	iso-Propyl Acetate	D	D	D_C
117	n-Butyl Acetate	D	D	D_C
118	sec-Butyl Acetate	D	D	D_C
119	iso-Butyl Acetate	D	D	D_C
120	n-Amyl Acetate	D	D	D_C
121	iso-Amyl Acetate	D	D	D_C
122	Methylamyl Acetate*	D	D-C	D_C
123	Vinyl Acetate	D	Training	
124	Methyl Acrylate*	D	D	D_C
125	Ethyl Acrylate	D	Training	
126	n-Butyl Acrylate*	D	C	D_C
127	iso-Butyl Acrylate*	D	D-C	D_C
128	2-Ethylhexyl Acrylate	D	C	D_C

TABLE VIII cont'd. TEST COMPOUNDS

Number	Compound Name	Classification NAS	MULTI	PLANE
129	iso-Decyl Acrylate	D	D	D_C
130	Methyl Methacrylate*	D	D	D_C
131	Propiolactone*	C	B	C_B
132	Caprolactone*	D	C	C_B
133	o-Dibutyl Phthalate	D	D	D_C
134	o-Diheptyl Phthalate*	D	C	D_C
135	Dioctyl Phthalate*	D	C	D_C
136	Dinonyl Phthalate*	D	C	D_C
137	Diisodecyl Phthalate*	D	C	D_C
138	Diundecyl Phthalate*	D	C	D_C
139	Butyl Benzyl Phthalate*	D	D	D_C
140	Ethyl Ether	C	Training	
141	iso-Propyl Ether	D	Training	
142	Ethylene Oxide	B	Training	
143	Propylene Oxide	B	Training	
144	Tetrahydrofuran	C	Training	
145	1,4 Dioxane	C	C	C_D
146	Morpholine	C	Training	
147	Epichlorohydrin	C	Training	
148	Dichloroethyl Ether	-	D	Unclassified**
149	Methyl Formal	C	Training	
150	Propyl Formal*	C	C	C_D
151	n-Butyl Formal*	C	C	C_D
152	iso- Butyl Formal*	C	C	C_D
153	Furfuryl Alcohol	C	D	D_C
154	Ethylene Glycol	D	C	C_D
155	Propylene Glycol	D	D	C_D
156	1,3 Butylene Glycol	D	C	D_C

TABLE VIII cont'd. TEST COMPOUNDS

Number	Compound Name	Classifications NAS	MULTI	PLANE
157	Hexylene Glycol	d	C	D_C
158	Ethylene Glycol Monomethyl Ether	C	C	C_D
159	Ethylene Glycol Monoethyl Ether	C	C	C_D
160	Ethylene Glycol Monobutyl Ether	C	C	C_D
161	Diethylene Glycol	C	C	C_B
162	Diethylene Glycol Monomethyl Ether*	C	C	C_D
163	Diethylene Glycol Monoethyl Ether*	C	C	C_D
164	Diethylene Glycol Monobutyl Ether*	C	C	C_D
165	Diethylene Glycol Monobutyl Ether Acetate	C	D	D_C
166	Dipropylene Glycol*	C	C	C_D
167	Triethylene Glycol	C	D	C_D
168	Tripropylene Glcyol*	C	C	Unclassified**
169	Methoxy Triglycol*	C	C	Unclassified**
170	Ethoxy Triglycol*	C	C	Unclassified**
171	Tetraethylene Glycol*	C	C	Unclassified**
172	Ethylene Glycol Monoethyl Ether Acetate	C	D	D_C
173	Ethylene Glycol Monobutyl Ether Acetate	C	D	D_C
174	Triethylene Glycol Di-(2-Ethyl Butyrate)*	C	C	D_C
175	Glycol Diacetate*	D	D	D_C
176	2-Hydroxyethyl Acrylate*	D	D-C	D_C
177	Glycerine	D	C	C_D
178	Methyl Chloride	D	D	D_C
179	Methylene Chloride	D	D	D_C
180	Methyl Bromide	D	D	D_C
181	Ethyl Chloride	D	D	D_C

TABLE VIII cont'd. TEST COMPOUNDS

Number	Compound Name	Classifications NAS	MULTI	PLANE
182	Ethylene Chloride	D	Training	
183	1,1,1-Trichloroethane	D	D	D_C
184	1,2-Dichloropropane	D	D	D_C
185	Ethylene Chlorohydrin	D	D	Unclassified**
186	Vinyl Chloride	D	Training	
187	Vinylidene Chloride	D	D	D_C
188	Trichloroethylene	D	D	D_C
189	Dichloropropane	D	D	D_C
190	Allyl Chloride	D	D	D_C
191	Chlorobenzene	D	D	D_C
192	o-Dicholorobenzene	D	D	Unclassified**
193	1,2,4-Trichlorobenzene	D	D	Unclassified**
194	Acetone	D	D	D_C
195	Methyl-Ethyl Ketone	D	D	D_C
196	Methyl Isobutyl Ketone	D	D	D_C
197	Diisobutyl Ketone*	D	D	D_C
198	Mesityl Oxide	D	Training	
199	Cyclohexanone	D	D	Unclassified**
200	Isophorone	D	D	D_C
201	Acetonitrile	D	D	D_C
202	Acrylonitrile	D	Training	
203	Ethylene Cyanohydrin	D	D	D_C
204	Acetone Cyanohydrin	D	D	D_C
205	Adiponitrile*	D	D	Unclassified**
206	Ethylene	C	Training	
207	Prppylene	C	Training	
208	Butylene	D	C	C_D

TABLE VIII cont'd. TEST COMPOUNDS

Number	Compound Names	NAS	Classifications MULTI	PLANE
209	Butadiene	B	Training	
210	1-Pentene	D	C	C_D
211	Isoprene	D	Training	
212	Hexene	D	B	C_D
213	Heptene	D	B	C_D
214	Octene	D	B	C_D
215	Diisobutylene	D	Training	
216	Nonene	D	C	C_D
217	Tripropylene*	D	C	C_D
218	Decene	D	C	C_D
219	Turpentine	D	D	D_C
220	Dipentene	D	C	C_D
221	Undecene*	D	C	D_C-C_D
222	Dodecene*	D	C	D_C
223	Tetrapropylene*	D	C	Unclassified**
224	Tridecene*	D	C	C_D
225	Tetradecene	D	C	C_D
226	Dicylcopentadiene*	C	C	C_D
227	Acetylene	A	Training	
228	Methyl Acetylene-Propadiene	B	-	-
229	Aluminum Triethyl	-	-	-
230	Ammonia (anhydrous)	D	Training	
231	Carbon Disulfide	A	Training	
232	Dimethylformamide	D	D	D_C
233	unsym-Dimethyl Hydrazine	C	Training	
234	Monomethyl Hydrozine*	C	C	B_C

TABLE VIII cont'd TEST COMPOUNDS

Numbers	Compound Name	Classifications NAS	MULTI	PLANE
235	2-Nitropropane	C	Training	
236	Nitrobenzene	D	C	C_D
237	Dinitrotoluene	C	C	Unclassified**
238	Hydrogen	B	Training	
239	Hydrogen Sulfide	C	Training	
240	Phenol	D	D	D_C

Compounds and NAS classifications are from "Matrix of Electrical and Fire Hazard Properties and Classfications of Chemicals" National Academy of Sciences, Washington, D. C. (DOT-CG-41680-A), 1975.

[a]MULTI is a multicategory separator contained in the statistical package routine called ARTHUR. (Appendix II)

[a]PLANE is a two category separator contained in the statistical package routine called ARTHUR. The subscript denotes the choice between the two categories the compound was not classified as. (Appendix II)

*These compounds had auto-ignition temperatures and/or solubilities missing. A range of their possible values was made by examining similar compounds and trends within the groups. A maximum interval of 50° was used for the auto-ignition temperature and of one unit for the solubilities. In cases in which a decision could not be made both chosen categories are listed.

**PLANE, which examines only two groups at a time, did not give a unique classification for all three pairs.

TABLE IX

Compounds Classified Upwards

Methyl Alcohol
Ethyl Alcohol
iso-Propyl Alcohol
n-Butyl Alcohol
n-Anmyl Alcohol
iso-Anmyl Alcohol
Hexanol
Methylamyl Alcohol
Methyl Isobutyl Alcohol
Ethyl Butanol
Cyclohexane
n-Octyl Alcohol
iso-Octyl Alcohol
2-Ethyl Hexanol
Nonyl Alcohol
Diisobutyl Carbinol
n-Decyl Alcohol
iso-Decyl Alcohol
Undecanol
Dodecanol
Tridecanol
Pentadecanol
iso-Pentyl Aldehyde
2-Ethyl Hexaldehyde
iso-Octyl Aldehyde
Glyoxal
n-Pentane
n-Hexane
n-Heptane
Octanes

Cyclohexane
Diethanolamine
Diisopropanolamine
n-Aminoethyl Ethanolamine
Diethylenetriamine
Triethylene Tetramine
Tetraethylene Pentamine
Decyl Benzene
Undecyl Benzene
Dodecyl Benzene
Tridecyl Benzene
Tetradecyl Benzene
Methylamyl Acetate
n-Butyl Acrylate
iso-Butyl Acrylate
2-Ethylhexyl Acrylate
Propiolactone
Caprolactone
O-Diheptyl Phthalate
Dioctyl Phthalate
Dinonyl Phthalate
Diisodecyl Phthalate
Diundecyl Phthalate
Ethylene Glycol
Propylene Glycol
1,3 Butylene Glycol
Hexylene Glycol
Triethylene Glycol

2 Hydroxyethyl Acrylate
Glycerine
Butylene
1-Pentene
Hexene
Heptene
Octene
Nonene
Tripropylene
Decene
Dipentene
Undecene
Dodecene
Tetrapropylene
Tridecene
Tetradecene
Momethyl Hydrazine
Nitrobenzene
Acrylic Acid
Nonane

FASCHING/STROMBERG
COAST GUARD DATA

CODE # 2163.1291

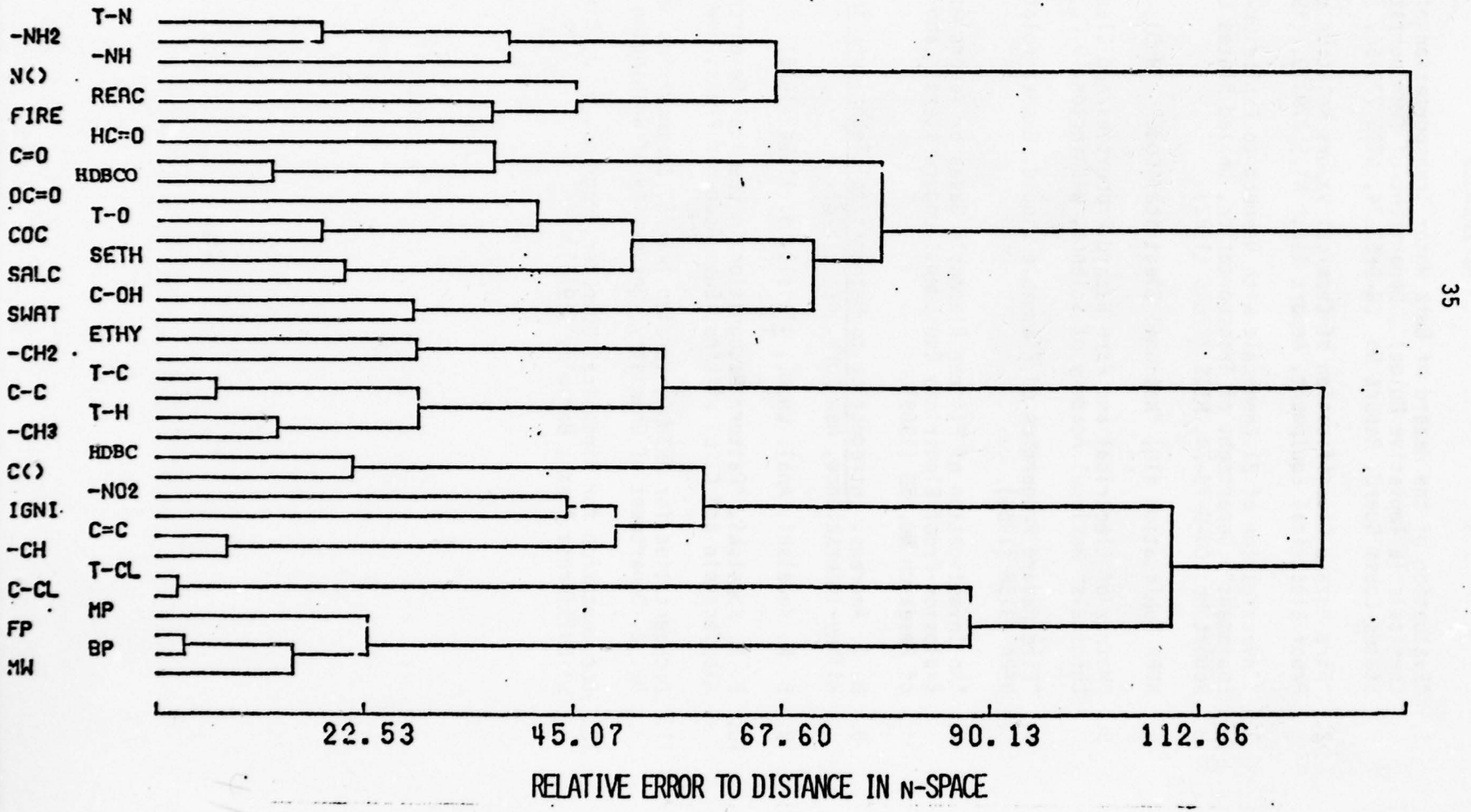


Figure 1

REFERENCES

1. "Evaluation of the Hazard of Bulk Water Transportation of Industrial Chemicals (a Tentative Guide)", Department of Transportation, United States Coast Guard, Report No. CG-D-63-74, NTIS 775756, (1973).

2. "Fire Hazard Classification of Chemical Vapors Relative to Explosion-Proof Electrical Equipment, Report III," NTIS 779179, (1973).

3. "Investigation of 21 Chemicals with Respect to Explosion-Proof Electrical Equipment", Department of Transportation, United States Coast Guard, Report No. CG-D-70-74, NTIS 777646 (1972).

4. NFPA Publication #10, "National Electrical Code" (1968).

5. "Matrix of Electrical and Fire Hazard Properties and Classification of Chemicals" National Academy of Science, Washington, D.C., NTIS A027181 (1975).

6. "Fire Hazard Properties of Flammable Liquid, Gases, Volatile Solids", NFPA #325M (1969).

7. "An Investigation of Fifteen Flammable Gases or Vapors with Respect to Explosion-Proof Electrical Equipment" Underwriters' Laboratories, Bulletin of Research No. 58 (1969).

8. H. C. Andrews, Introduction to Mathematical Techniques in Pattern Recogniton, Wilkey-Interscience, New York, NY (1972).

9. B. R. Kowalski, Anal. Chem., 47, #13, p. 1152a (1975).

10. B. R. Kowalski, "Pattern Recognition in Chemical Research", Vol. 2, C.E. Klopfenstein and C. L. Wilkins, Eds. Academic Press, New York, NY (1974).

11. Documentation for ARTHUR, Version 1-8-75, Chemometrics Society Report No. 2, Department of Chemistry, University of Washington (1975).

12. Documentation for Biomedical Computer Programs, W. J. Dixon, ed., University of California Press, Berkeley (1975).

APPENDIX I

Simple Experiments for Understanding

Factor Analysis and Hierarchial Clustering

INTRODUCTION

Modern analytical tools, such as neutron activation analysis and atomic absorption spectroscopy, have enabled scientists to collect large amounts of quantitatively accurate information from individual samples. When many samples are involved, the scientist is then faced with the dilemma of interpreting his data. The conventional first step is to place all of the data in table form. Examining multivariable data tables in this way can cause eye strain, but, except where data values are unusually different, it can often lead to little else. Simple statistics such as standard deviations and t-tests may tell the scientitst which are outliers, but once again will often show him little of the complex interrelationships among the variables or samples. The researcher may then plot two-dimensionally certain variables of his data versus other variables. This step can be a great aid to interpretation since he can now see a spacial representation of relationships among the selected variables. At the same time, however, it is quite limited in the amount of information that can be displayed.

Another step which has recently been applied to chemical problems is computerized pattern recognition, in which all of the variables (or samples) may be compared to one another to determine their inter and intra-relationships. Pattern recognition is a developing branch of artificial intelligence (1) which has been used for such diverse purposes as medical diagnosis (2,3), the identification of rocks (4), and hand drawn character identification (5). Jurs (6), Kowalski (7), and Isenhour(8) have described how pattern recognition can be useful in solving a variety of chemical problems. Chemical applications have included the identification and interpretation of mass spectra data (9), IR specta (10), NMR data (11),

gamma-ray spectroscopy from neutron activation (12), and stationary electrode polarography (13). Other studies have included the determination of the correct chlorine dosages for water treatment (14); the relationship between mass spectra data and the pharmacological activity of drugs (15); analysis for oil in natural waters (16); screening prospective anti-cancer drugs (17); and the classification of archeological artifacts from trace element data (18).

Factor analysis is a form of pattern recognition in which the linear combinations of a set of experimental data are developed and this hopefully reduces the number of variables. Its method has been described in detail by Veldman (19) and Harman (20). This technique has been applied to such diverse areas as biology to determine the growth patterns in plants (21); psychology to study word recognition (22) and cultural differences (23); meteorology to study coastal air and water temperatures (24); and geology to define deformational modes in rock (25). Chemists have used factor analysis to study data from nuclear magnetic resonance spectroscopy (26) and from gas-liquid chromatography (27). Factor analysis has also been used to correlate trace element and other chemical data collected from a number of samples. Examples include the study of chemical pollutants in air samples (28, 29) and the correlation of rocks based on their chemical composition (30, 31).

Pre-treatment of the raw data may include normalizing the variable (or sample) values to the mean standard deviation. The data may then be reduced to a correlation coefficient matrix. A number of correlation coefficient methods may be used, including cosine coefficient, distance coefficient, and Horner coefficient. The product moment correlation coefficient is used in this article's examples:

$$r_{xy} = \frac{\Sigma(x_i-\bar{x})\ (y_i-\bar{y})}{[\Sigma(x_i-\bar{x})^2]^{1/2}[\Sigma(y_i-\bar{y})^2]^{1/2}}$$

The factor analysis method used by the authors takes this correlation coefficient matrix and determines the set of eigenvalues for the linear combinations, the cumulative percentages of these eigenvalues, the eigenvectors, and finally the loaded factor matrix for each of the eigenvalues. The general method model used, that of principal components was:

$$z_j = a_{j1}F_1 + a_{j2}F_2 + \ldots + a_{jn}F_n \quad (j=1,2, \ldots,n)$$

where each of the n observed variables of the new data matrix was described linearly in terms of the new uncorrelated components, F (20). The "a" coefficients are the factor loadings. Those eigenvalues considered to be significant factors are retained: significance usually being defined as a value greater than or equal to 1.00. The loaded factor matrix of significant factors then undergoes varimax rotation in order to maximize the differences among the factors. This rotated factor matrix is normalized to range from -1.0 to +1.0. A positive value of a variable in a factor shows a direct relationship of the variable to that factor. The greater the value the stronger the relationship indicated. A negative number shows some inverse or negative relationship and a value close to zero infers that there is no direct relationship between variable and factor. This rotated factor matrix may then be studied in either table or graphic form in order to interpret the initial data.

Another useful pattern recognition method is hierarchial clustering (19). This unsupervised learning method clusters the samples from either

the raw data or a normalized data matrix according to their n'th dimensional distance vector across the variables. The mean distance between samples or clusters is used to determine the relative error of the grouping and used as the new vector distance value for future groupings. Those groups closest in distance values will cluster first. Eventually, all groups will be clustered into two groups. A dendrogram can then be made of the series of clusters to give a graphical representation of the calculations. The original data matrix may be transposed and similar clustering may be made of the variables as they vary across the samples.

EXPERIMENTAL

The authors have developed a FORTRAN IV computer program to handle statistical evaluation of data, perform correlation analysis, factor analysis and hierarchial clustering, and to display data and results in either table or graphical form (Table I). All calculations were done at the University of Rhode Island's Computer Science Center on an IBM-370/155 computer and graphics were done on a Broomall Industries 2000 Series Incremental Plotter. Data input to the programs is accepted from either cards or from general disk storage data banks. Four correlation coefficients are presently available: product moment correlation coefficient, cosine coefficient, distance coefficient, and Horner coefficient. The graphic displays can handle any data matrix from the routines, from the raw data to any calculated coefficients. Almost all routines may be accessed at any time during program operation, specific use being governed by program read control cards. Once the raw data has been entered, it may be treated by any of the procedures and the output may be returned to the user in either table or graphic form. The programming package has been designed so that the user need not have programming

experience to operate it.

To illustrate factor analysis as an interpretive tool, a synthetic data set consisting of fifty groups of length, width and height values, or in other words, fifty random boxes, was generated from a random number table (32).

A group of linearly related variables (Table II) were generated from the length, width and height values. Using factor analysis these ten variables were reduced to three significant factors, each containing about one third of the total variation among the variables. The rotated factor matrix is shown in Table III. It is useful to graph the variable values of the rotated factor matrix as they vary across the factors. The two dimensional plot of factor one versus factor two (Figure 1a) indicated that the length variable was strongly associated with factor one while unrelated to factors two and three. The width variables was strongly associated with factor two and unrelated to factors one and three, and the height variable was unrelated to either factor one or two. The linear combination variables were arranged according to their weighted length, width or height value. The plot of factor one versus factor three (Figure 1b) was nearly identical to the previous figure, except that width and height has been reversed. Plotting factor two versus factor three also showed a similar result (Figure 1c), this time reversing length and width. Since each factor contained about 33 percent of the total variation, each two-dimensional plot could only give about 2/3 of the information available. Comparison of these three factors on one three-dimensional plot (Figure 2) simplified the interpretation of the problem by allowing 100 percent of the information to be presented at one time. The x-axis (right side) and y-axis (left side) represented

factors one and two, respectively, and factor three was the z-axis. The peaks in the rear three corners represented a value of 0.0, or in other words, a non-relationship of the variable to factor three. It was interesting to note in Figure 2 the successive progression along diagonals of the associated length-weighted, width-weighted, and height-weighted variables. The valid interpretation from this information was that factors one, two and three were actually length, width and height, respectively. Such an interpretation would be nearly impossible to make from observation of the raw data alone. Table IV is a partial listing of the initial input information for this example.

Factor analysis was designed to associate linear related variables, but it may also be used to correlate variables with non-linear relationships. To prove this point, a variable data matrix of cross product information from the boxes (Table II) was tested in a similar manner. When these variables were handles exactly the same as the preceding examples, three nearly equal factors were again obtained from factor analysis, although in this case they contained about 92 percent of the variation instead of the 100 percent found in the linear example. When their relative positions on the three-dimensional plot of these three significant factors were observed (Figure 3), the variables length, width and height were very strong in factors one, two and three respectively. The interpretation once again was that length, width and height were the three significant factors, as would be expected.

As an added test of factor analysis, the linear and cross-product data sets were then combined and tested the same way. Again three significant factors were found, this time accounting for 96 percent of the variation. There were no significant differences in the rotated

factor matrix values from the previous two determinations, and a study of the three-dimensional representation of the three factors (Figure 4) showed nearly the same result that could be found if the three-dimensional plots of the two data sets alone were superimposed.

Two additional tests were necessary to confirm the validity of factor analysis in circumstances where the answer is not known before hand. Subgroups of 40, 30,20 and 10 boxes from the linear variable data set were studied (Table V) to determine the effect of sample size on the results. In the second test, the values for the variables length, width and height were deleted from the data matrix before factor analysis in order to determine if the use of these three variables was biasing the results. No significant differences were found in either experiment from those results in the initial studies. Caution should be taken in applying these results when interpreting real as opposed to synthetic data. The size of the sample set is important, too few samples can cause an incorrect clustering and hence false interpretations of the data. A minimum of at least twice as many samples as variables is necessary.

It is possible with this program package to rotate the three-dimensional representation about the z-axis or in the X-Y plane. The best view is data dependent because cluster representations can mask each other. The three-dimensional representation of the linear box variable factor matrix was used in Figure 5 to demonstrate this rotation. Figures 5a and 5b show the plot rotated to relative positions of 20° and 70° about the z-axis while maintaining the X-Y plane at 45°. In Figures 5c and 5d, the z-axis position has been returned to 45° and the X-Y plane rotated to 20° and 70° respectively.

Hierarchial clustering was applied to the same boxes and their associated variables, which were examined earlier using factor analysis. The

dendrogram of the linear variables of the boxes (Figure 6) showed a clear separation of three cluster sets: length with length weighted variables, width with width weighted variables, and height with height weighted variables. As in factor analysis test, this was the relationship that would be expected to occur among variables which are known to have a linear relationship to one another. The same clustering method was then used on the cross-product variables (Figure 7). The interpretation of this plot was less defined than the first example. Length clustered with length side diagonal and area information, and width clustered with width side diagonal and area information. Total volume and total surface area also clustered with one another. Since hierarchial clustering is an unsupervised learning method, however, the multi-interrelationships among a set of cross-product related variables tend toward noninterpretive clustering by this method. When both sample sets were combined and tested by hierarchial clustering, the resulting dendrogram (Figure 8) showed properties similar to each of the previous two figures, that is, the linear variables were clustered into three readily apparent groups of length-weighted, width-weighted, and height-weighted values, and the two variables of total diagonal and length-plus-width-plus-height variables also clustered closely.

One final experiment was performed on these boxes. The data matrix was transposed and the fifty boxes themselves were compared to one another as they varied across the linear relationship variables. Factor analysis gave three significant factors, each with about one-third of the total information. Hierarchial clustering also showed three distinctly separate clusters (Figure 9), which can be accounted for by the general groupings of boxes with a large width values and usually large height value

bias, boxes with a small width bias with neither a length or height value bias, and boxes with a small length value and a small height value bias with no bias of the width value.

LITERATURE CITED

(1) Nilsson, N. J., "Learning Machines", McGraw Hill Book Co., New York, 1965.

(2) Winkel, P., Clinical Chemistry, 19(12), 1329 (1973).

(3) Patrick, E. A., Stelmack, F. P., and Shen, L. Y., IEEE Transactions on Systems, Man, and Cybernetics, SMC-1, 1 (1974).

(4) Conley, C. D., and Davis, J. C., Am. Assn. Pet. Geol. Bull., 57, 399 (1973).

(5) Sklansky, J., "Pattern Recognition: Introduction and Foundations", Dowden, Hutchinson and Ross, Inc., Stoudsburg, PA, 1973, p. 328.

(6) Jurs, P. C., Kowalski, B. R., Isenhour, T. L., and Reilley, C. N., Anal. Chem., 41, 1949 (1969).

(7) Kowalski, B. R., and Bender, C. F., J. Amer. Chem. Soc. 94, 5632 (1972).

(8) Isenhour, T. L., and Jurs, P. C., Anal. Chem., 43, 20a (1971).

(9) Felty, W., and Jurs, P. C., Anal. Chem., 45, 885 (1973).

(10) Kowalski, B. R., Jurs, P. C., Isenhour, T. L., and Reilly, C. N., Anal. Chem., 41, 1945 (1969).

(11) Kowalski, B. R., and Reilley, C. A., J. Phys. Chem., 75, 1402 (1971).

(12) Wangen, L. E., and Isenhour, T. L., Anal. Chem., 42, 737 (1970).

(13) Sybrandt, L. and Perone, S., Anal. Chem., 44, 2331 (1972).

(14) Kuo, K. S. and Jurs, P. C., Am. Water Works Assn. J., 65, 623 (1973).

(15) Ting, K. L., Science, 180, 417 (1973).

(16) Lysyj, I., and Newton, P. R., Anal. Chem., 44, 2385 (1972).

(17) Kowalski, B. R. and Bender, C. F., J. Am. Chem. Soc., 96, 916 (1974).

(18) Kowalski, B. R., Schatzki, T. F., and Stross, F. H., Anal. Chem., 44, 2176 (1972).

(19) Veldman, D. J., "Fortran Programming for the Behavior Sciences", Holt, Rinehart and Winston, New York, 1967.

(20) Harman, H. H., "Modern Factor Analysis", The University of Chicago Press, Chicago, 1970.

(21) Andel, J. V., Nelissne, H. J., Act. Bot. NEE, 22, 599 (1973).

(22) Hermann, D. J., Chaffin, R. J., and Corbett, A. T., Journal of Verbal Learning and Verbal Behavior, 12, 666 (1973).

(23) Gault, U., and Wang, A. M., Br. J. Soc. and Clin. Psychol., 13, 37 (1974).

(24) Berell, C. E., and Bundgaard, R., J. Appl. Meteorology, 10, 803 (1971).

(25) Tobin, D. C., and Donath, F. A., Geological Society of American Bulletin, 82, 1463 (1971).

(26) Weiner, P. H., Malinowski, E. R. and Levinstone, A. R., J. Phys. Chem., 74, 4537 (1970).

(27) Weiner, P. H. and Howery, D. G., Anal. Chem, 44, 1189 (1972).

(28) Blifford, I. H., and Howery, D. G., Anal. Chem., 44, 1189 (1972).

(29) John, W., Rahn, K., Kaifer, R., and Wesolowski, J. J., Atmospheric Environment, 7, 107 (1973).

(30) Dawson, K. M. and Sinclair A. J., Econ. Geol., 69, 406, (1974).

(31) Reeves, M. J. and Saadi, T. A., Econ. Geol., 69, 406 (1974).

(32) Weast, R. C., editor in chief, "CRC Handbook of Tables for Mathematics Forth Edition", The Chemical Rubber Company, Cleveland, p. 975 (1970).

(33) Zoller, W. H., Gladney, E. S., and Duce, R. A., Science, 193, 198 (1974).

LIST OF TABLES

LIST OF FIGURES

TABLE I

PROGRAM ROUTINES

Input	Statistics	Correlation Coefficients
Card	Arithmetic Mean	Cosine Coefficient
Disk	Geometric Mean	Distance Coefficient
	Median	Horner Coefficient
	Standard Deviation	Product Moment Coefficient
	Mean Std. Deviation	
	Second Moment	
	Third Moment	
	Skewness	
	Kurtosis	
	Missing Values	

Clustering	Graphics
Factor Analysis	2-Dimensional Line Printer
Hierarchial Clustering	2-Dimensional Computer Graphics
	3-Dimensional Computer Graphics
	Dendrogram Computer Graphics

TABLE II

Random Box Variables

Linear Sum Variables	Cross Product Variables
Length	Length
Width	Width
Height	Height
L+W+H	L W Diagonal
2L+W+H	L H Diagonal
L+2W+H	W H Diagonal
L+W+2H	Total Diagonal
3L+W+H	L W Rectangle Area
L+3W+H	L H Rectangle Area
L+W+3H	W H Rectangle Area
	Total Surface Area
	Volume

TABLE III

Rotated Factor Matrix
for Linear Box Variables

Variable	Factor 1	2	3
Length	0.9986	-	-
Width	-	0.9992	-
Height	-	-	0.9986
L+W+H	0.5835	0.5640	0.5842
2L+W+H	0.8172	-	-
L+2W+H	-	0.8146	-
L+W+2H	-	-	0.8065
3L+W+H	0.9031	-	-
L+3W+H	-	0.9071	-
L+W+3H	-	-	0.8915
% of Total Information	33.7 %	32.8 %	33.5 %

TABLE IV

Linear Variables for the First Ten Boxes

Box	Length	Width	Height	L+W+H	2L+W+H	L+2W+H	L+W+2H	3L+W+H	L+3W+H	L+W+3H
1	283.	517.	245.	1045.	1328.	1562.	1290.	1611.	2079.	1535.
2	732.	1.	838.	1571.	2303.	1572.	2409.	3035.	1573.	3247.
3	123.	596.	387.	1106.	1229.	1702.	1493.	1352.	2298.	1880.
4	281.	377.	796.	1454.	1735.	1831.	2250.	2016.	2208.	3046.
5	179.	136.	466.	781.	960.	917.	1247.	1139.	1053.	1713.
6	522.	400.	577.	1499.	2021.	1899.	2076.	2543.	2299.	2653.
7	229.	305.	859.	1393.	1622.	1698.	2252.	1851.	2003.	3111.
8	25.	484.	936.	1445.	1470.	1929.	2381.	1495.	2413.	3317.
9	584.	620.	505.	1709.	2293.	2329.	2214.	2877.	2949.	2719.
10	349.	777.	517.	1643.	1992.	2420.	2160.	2341.	3197.	2677.

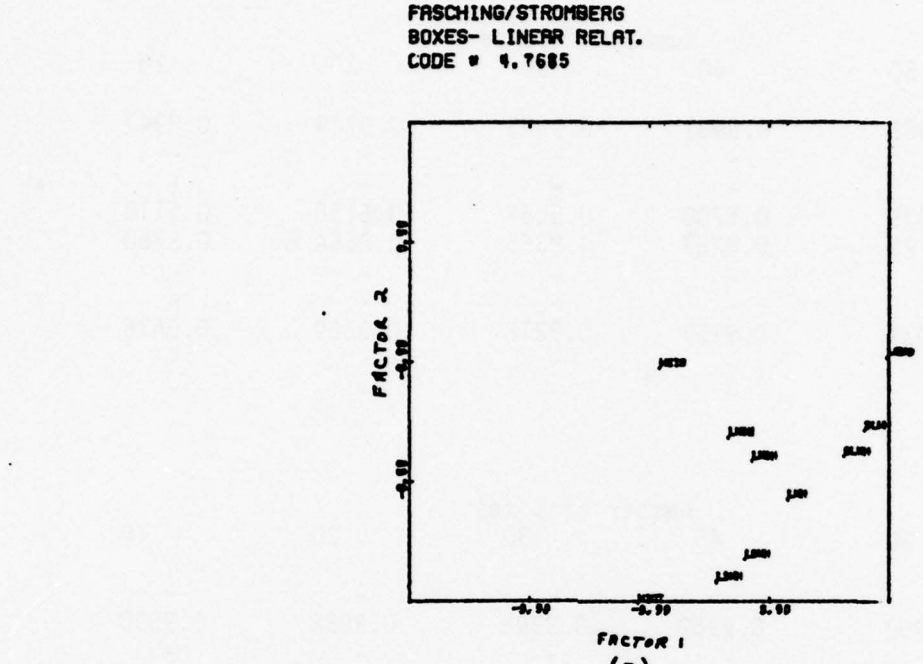
TABLE V

Rotated Factor Matrices of Different Sized Data Sets

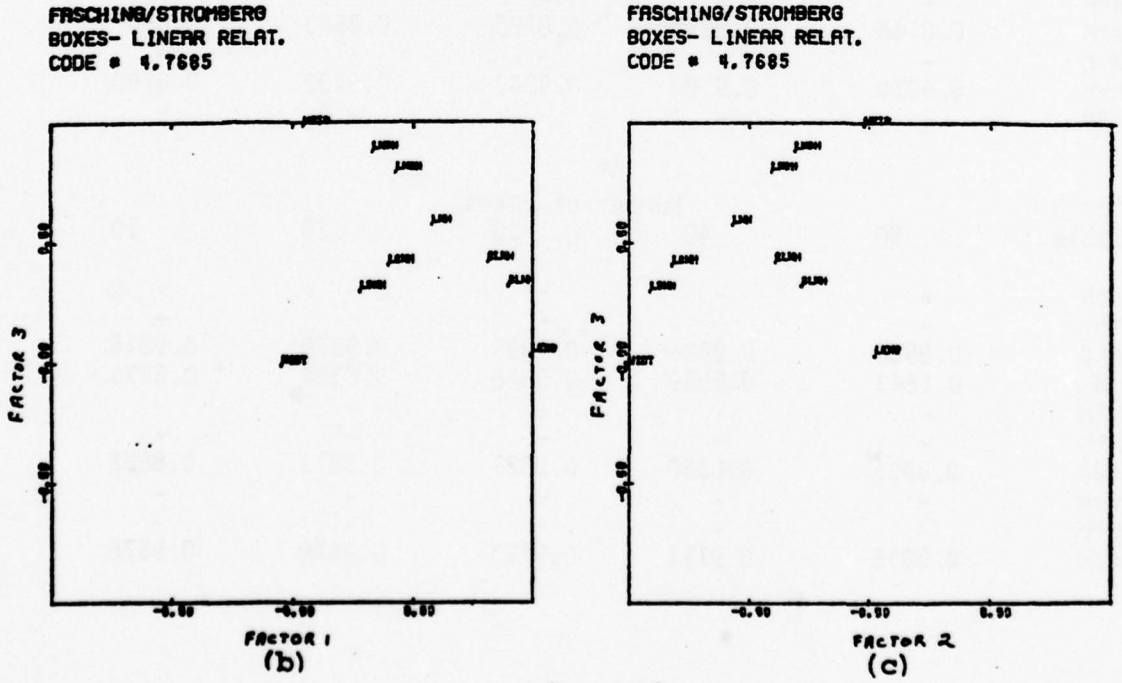
Factor 1	Number of Boxes				
Variable	50	40	30	20	10
Length	0.9986	0.9981	0.9993	0.9749	0.9947
Width	-	-	-	-	-
Height	-	-	-	-	-
L+W+H	0.5835	0.5780	0.5857	0.5158	0.6118
2L+W+H	0.8172	0.8287	0.8365	0.8564	0.8760
L+2W+H	-	-	-	-	-
L+W+3H	-	-	-	-	-
3L+W+H	0.9031	0.9157	0.9218	0.9559	0.9476
L+3W+H	-	-	-	-	-
L+W+3H	-	-	-	-	-

Factor 2	Number of Boxes				
Variable	50	40	30	20	10
Length	-	-	-	-	-
Width	0.9992	0.9969	0.9998	0.9898	0.9590
Height	-	-	-	-	-
L+W+H	0.5640	0.5532	0.5610	0.5713	0.5277
2L+W+H	-	-	-	-	-
L+2W+H	0.8146	0.8237	0.8120	0.8593	0.8966
3L+W+H	-	-	-	-	-
L+3W+H	0.9070	0.9190	0.9047	0.9432	0.9790
L+W+3H	-	-	-	-	-

Factor 3	Number of Boxes				
Variable	50	40	30	20	10
Length	-	-	-	-	-
Width	-	-	-	-	-
Height	0.9986	0.9996	0.9995	0.9979	0.9815
L+W+H	0.5843	0.5989	0.5848	0.6384	0.5893
2L+W+H	-	-	-	-	-
L+2W+H	-	-	-	-	-
L+W+2H	0.8065	0.8300	0.8321	0.8813	0.8852
3L+W+H	-	-	-	-	-
L+3W+H	-	-	-	-	-
L+W+3H	0.8915	0.9111	0.9177	0.9478	0.9576


FASCHING/STROMBERG
BOXES- LINEAR RELAT.
CODE # 4.7685
FACTOR 2
FACTOR 1

(a)


FASCHING/STROMBERG
BOXES- LINEAR RELAT.
CODE # 4.7685
FACTOR 3
FACTOR 1
FASCHING/STROMBERG
BOXES- LINEAR RELAT.
CODE # 4.7685
FACTOR 3
FACTOR 2

(b) (c)

Figure 1

FASCHING/STROMBERG
BOXES- LINEAR RELAT.
CODE # 4.7685

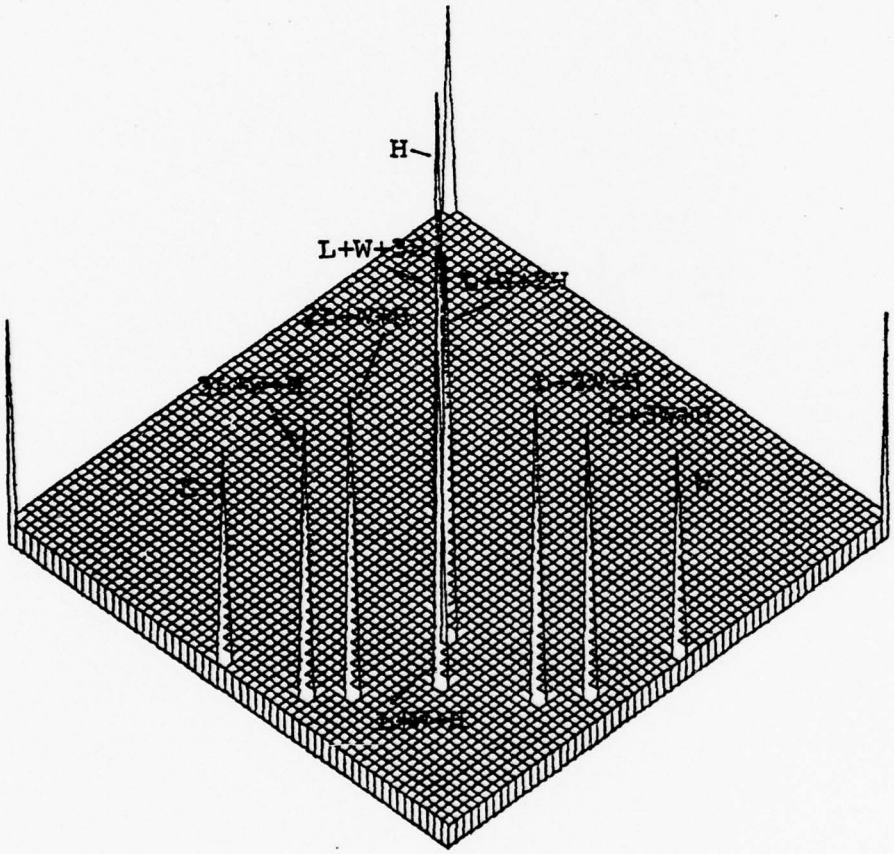


Figure 2

FASCHING/STROMBERG
BOXES- CROSS PROD.
CODE # 6.0955
45/45

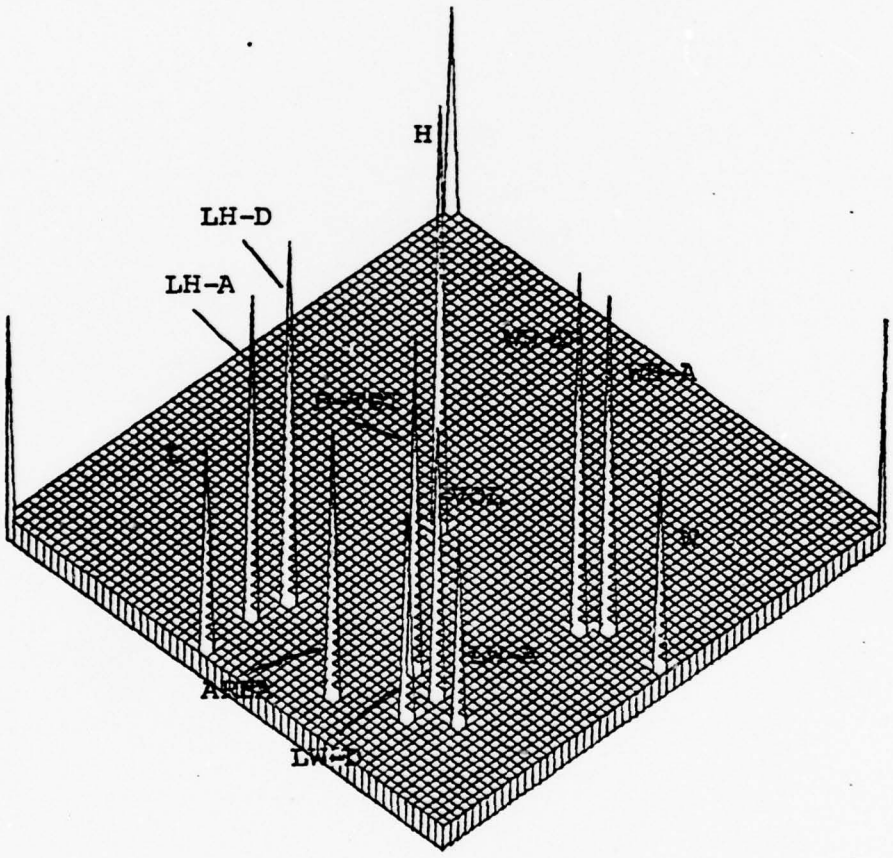


Figure 3

FASCHING/STROMBERG
50 RANDOM BOXES (3)
KZCOR
NORMALIZED BY COL.
CODE # 11.29757

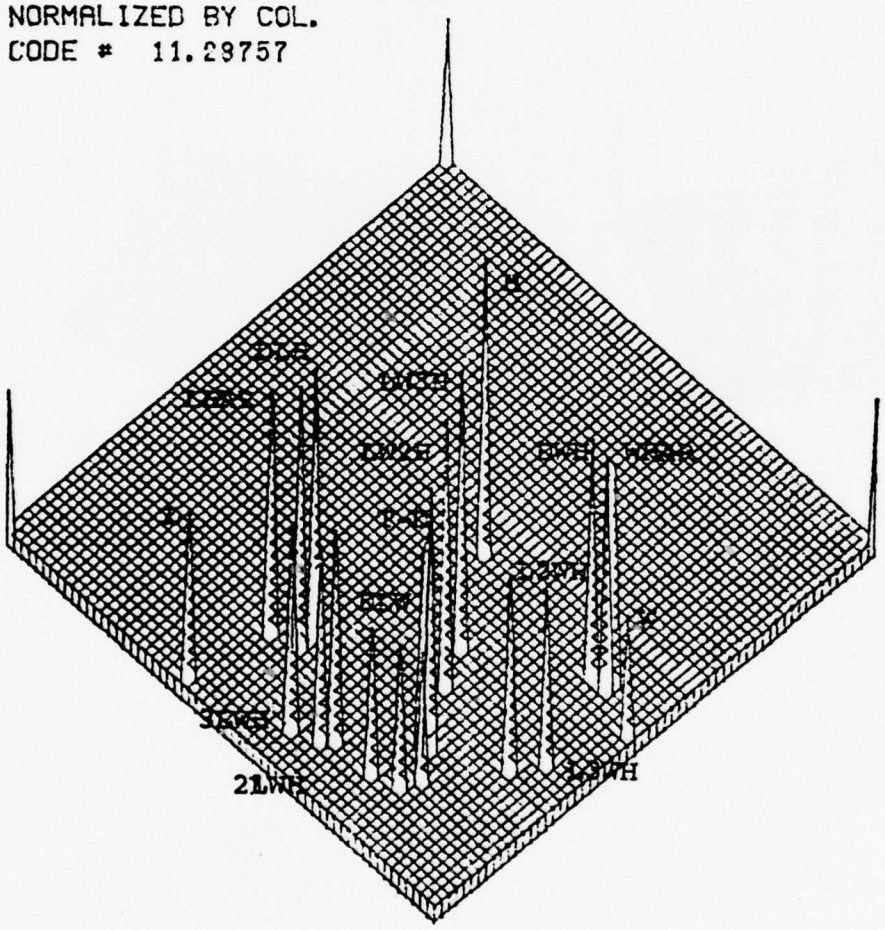
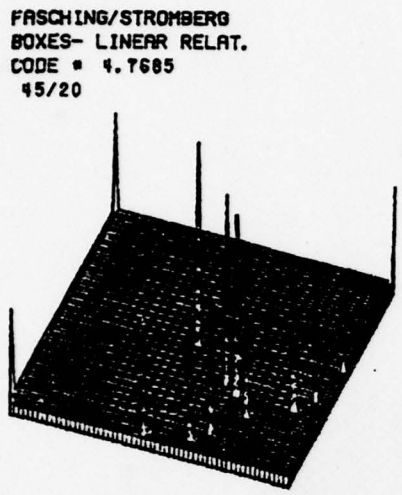
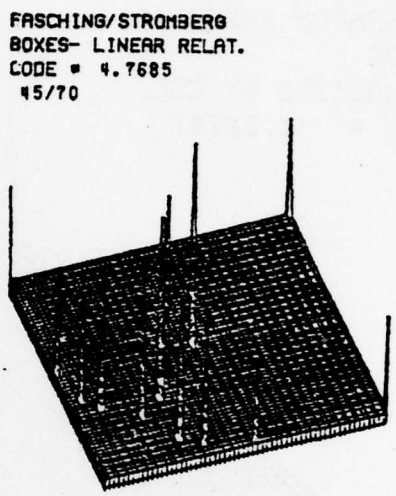


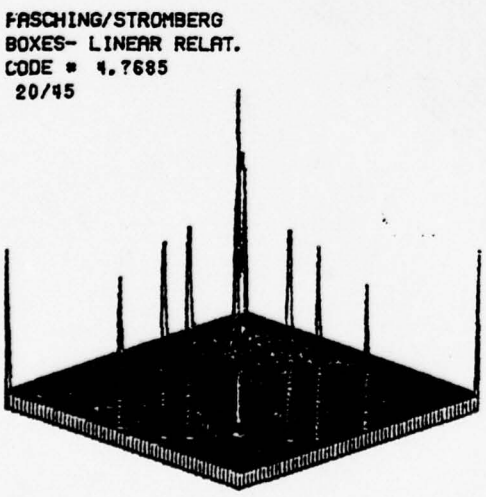
Figure 4



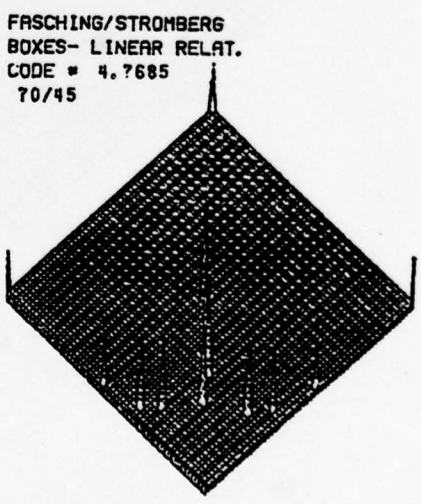
(a)



(b)



(c)



(d)

Figure 5

Figure 6

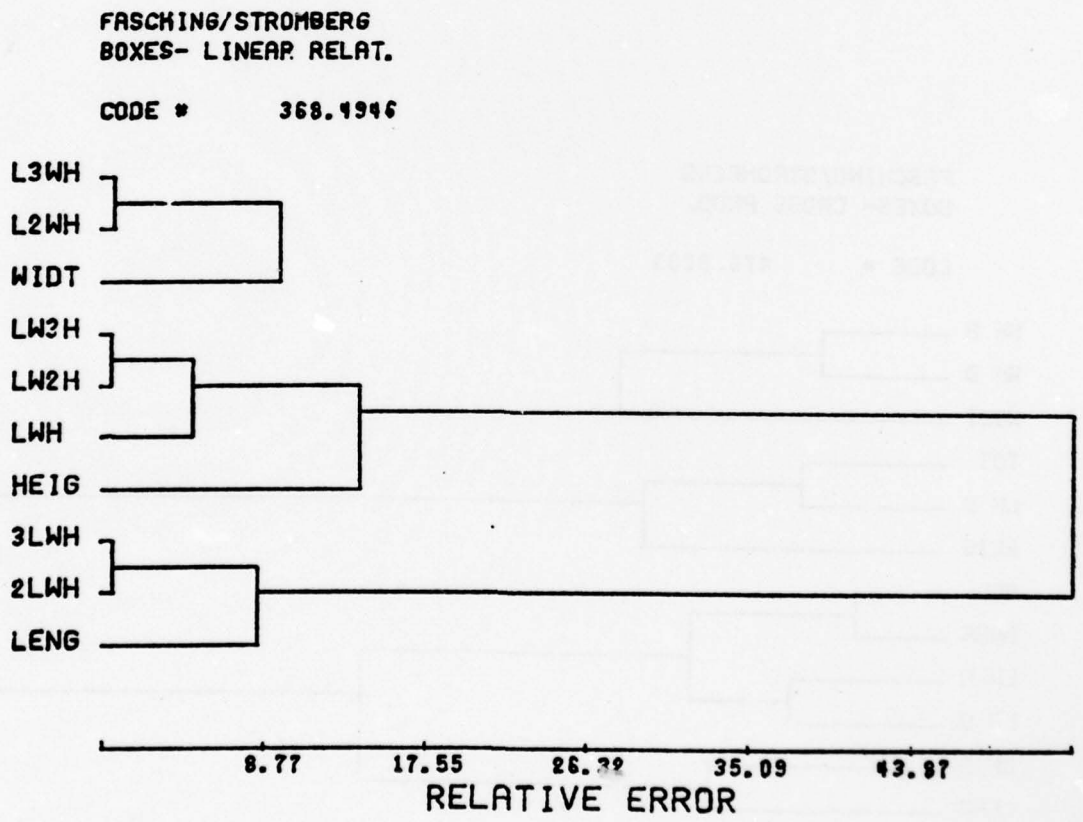

FASCHING/STROMBERG
BOXES- LINEAR RELAT.
CODE # 368.1946
L3WH
L2WH
WIDT
LW3H
LW2H
LWH
HEIG
3LWH
2LWH
LENG
8.77
17.55
26.32
35.09
43.87
RELATIVE ERROR

Figure 7

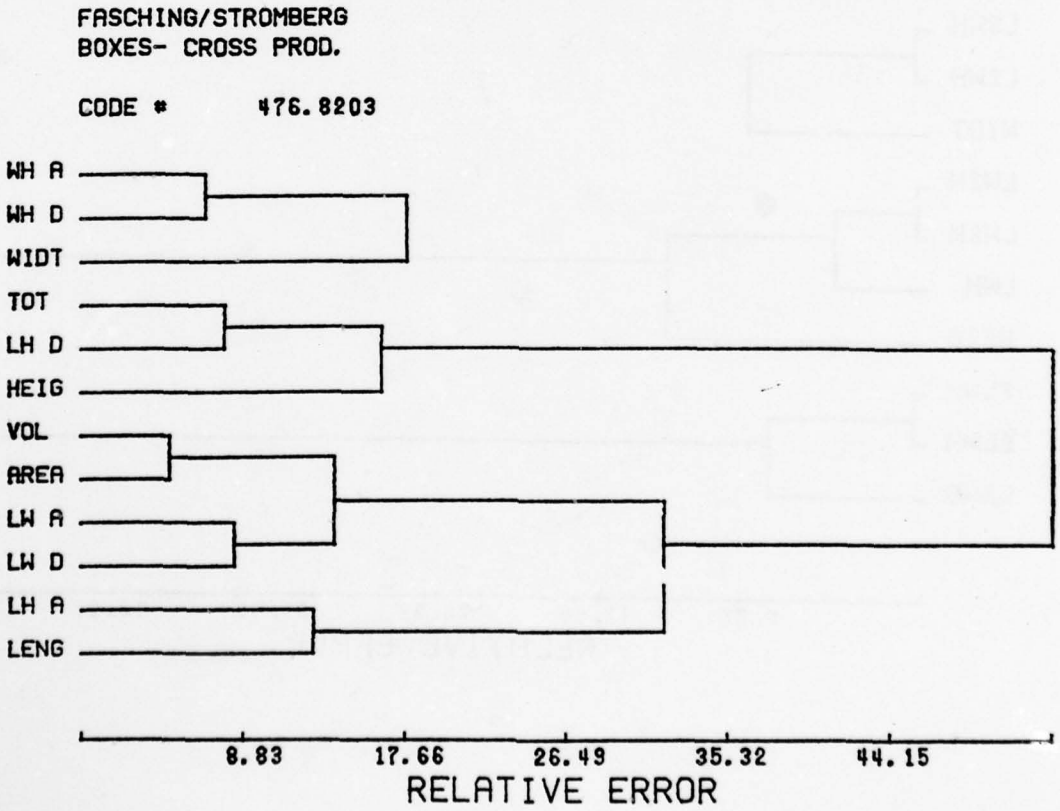

FASCHING/STROMBERG
BOXES- CROSS PROD.
CODE #
476.8203
WH A
WH D
WIDT
TOT
LH D
HEIG
VOL
AREA
LW A
LW D
LH A
LENG
8.83
17.66
26.49
35.32
44.15
RELATIVE ERROR

Figure 8

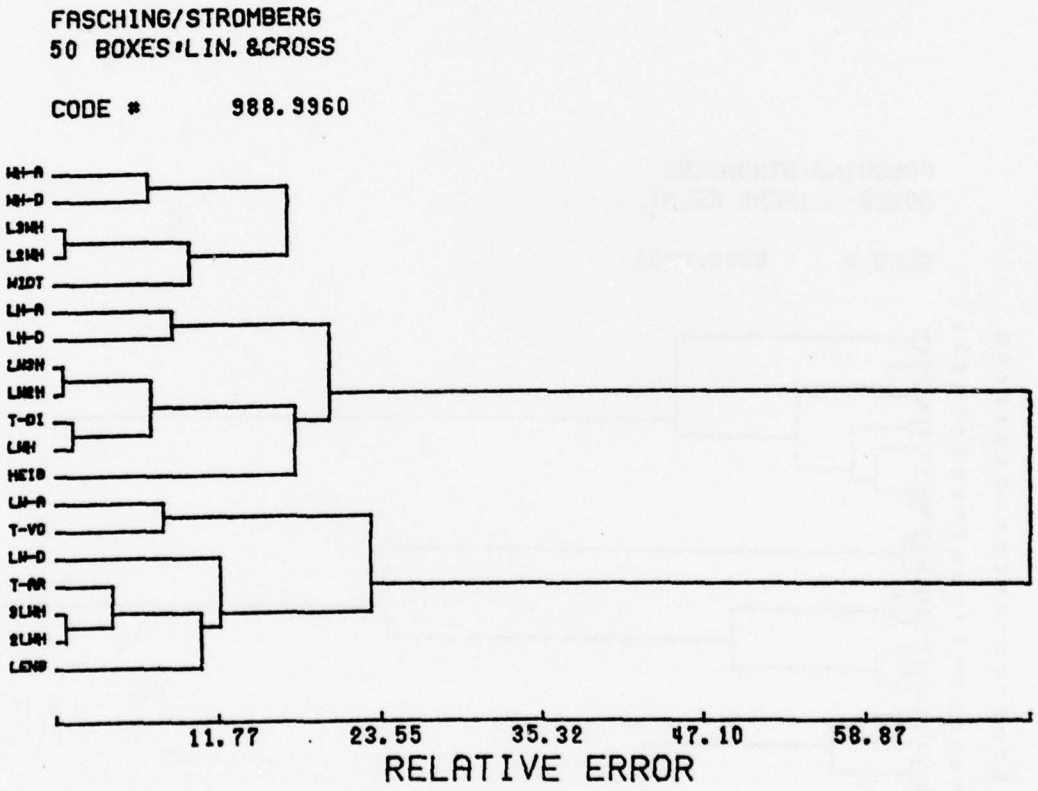
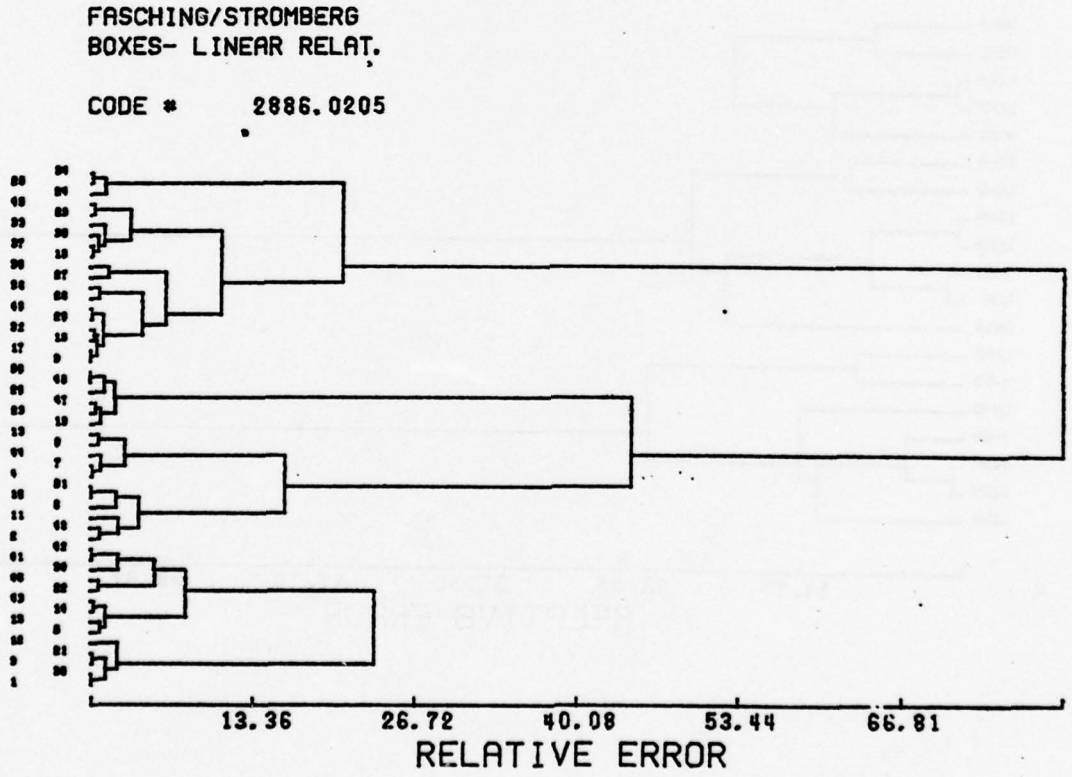

FASCHING/STROMBERG
50 BOXES*LIN. &CROSS
CODE * 988.9960
MH-A
MH-D
L3MH
L2MH
MIDT
LH-A
LH-D
LM3H
LM2H
T-DI
LMH
MEID
LH-A
T-VD
LH-D
T-AR
3LMH
2LMH
LEMH
11.77
23.55
35.32
47.10
58.87
RELATIVE ERROR

Figure 9


FASCHING/STROMBERG
BOXES- LINEAR RELAT.
CODE # 2886.0205
13.36
26.72
40.08
53.44
66.81
RELATIVE ERROR

APPENDIX II

A Description of ARTHUR

This appendix is abstracted from a paper entitled:

"ARTHUR and Experimental Data Analysis:

The Heuristic Use of a Polyalgorithm"

A. M. Harper, D. L. Duewer* and B. R. Kowalski
Laboratory for Chemometrics
Department of Chemistry
University of Washington
Seattle, Washington 98195

and

James L. Fasching
Department of Chemistry
University of Rhode Island
Kingston, Rhode Island 02881

"ARTHUR and Experimental Data Analysis: The Heuristic Use of a Polyalgorithm"

A. M. Harper, D. L. Duewer* and B. R. Kowalski
Laboratory for Chemometrics
Department of Chemistry
University of Washington
Seattle, Washington 98195

and

James L. Fasching
Department of Chemistry
University of Rhode Island
Kingston, Rhode Island 02881

Most non-routine data analysis in chemistry is designed to aid the formulation and/or evaluation of some model or hypothesis of the instrinsic data structure. The more detailed the model of the data's structure, that is, the more complete the analyst's understanding of the data, the more facile the selection of appropriate algorithms for the data analysis. Conversely, where very little is known of the data's structure it is difficult to make a priori selection or evaluation of analysis methodologies.

ARTHUR (1,2), a system of data manipulation, pattern recognition and robust statistical algorithms, is designed as a tool for the analyst in applications where the data's structure is not well understood. The algorithms included in the system are those which our laboratory and other members of the Chemometrics Society have found useful in the analysis of a number of quite different chemical and biological data sets. Recently implemented provisions for the inclusion of measurement uncertainties in the mathematical methods (3) enable the determination of which aspects of the data structure are truly inherent to the data. Descriptions of these algorithms can be found in the appendix of this chapter. It should be noted that ARTHUR is meant to be complementary to and not in competition with such primary statistical systems as SPSS (4) and BMD (5).

The primary utility of ARTHUR being in the formulation and evaluation of models for incompletely-understood data sets, it is not possible to specify given algorithms or sequences of algorithms which are "best". However, in the course of much data analysis (both fruitful and frustrating) some "rules of thumb" or heuristic procedures have been formulated. Following an introduction to the "ARTHURian" terminology of data analysis and pattern recognition, and a description of the inclusion of measurement uncertainties in pattern recognition methods, the heuristic techniques the developers and users of ARTHUR have found most generally useful will be described.

Definitions

The following terms and definitions have proved useful in describing the types of data analysis algorithms available in ARTHUR and in describing the data to be analyzed.

Classification Analysis

The data are known to be composed of specified groupings or categories. The goals of such analysis are the identification of what parameters (if any) qualitatively distinguish the known groupings and (if possible) the selection of a classification rule for identifying the known groups.

Continuous Property Analysis

The data are known to represent a continuous range of responses towards some given property(ies). The goals of such analysis are the identification of what parameters (if any) are functionally related to the property and (if possible) the selection of a rule which quantitatively predicts that property.

Unsupervised Analysis

The data are not known to have any systematic characteristics. The goal of such analysis is the discovery of what systematic behavior the data exhibit (if any exists). Study of the regularities among objects is generally referred to as cluster analysis; study of the regularities among measurements is generally referred to as factor analysis.

Object

A compound, sample, individual or other entity for which a list of characterizing parameters is present in the data base.

Measurement

An experimentally available parameter (independent variable) used to characterize the objects.

Feature

Any transformation of one or more measurements used to characterize the objects. When referring to a parameter which can be either a measurement or a feature, the term "measurement/feature" is used.

Data vector

The complete list of measurement/feature values used to characterize a particular object. (The older chemical pattern recognition literature, including that of the Laboratory for Chemometrics, refers to this as a "pattern". Considerable semantic confusion over "patterns of patterns" forced the change to the term, "data vector".)

Category

One of the groups of objects studied in the classification analysis algorithms. Categories which are entirely independent of one another, such as the labeling of white bond papers by their manufacturer, are referred to as discrete categories. Categories which have some dependence upon one another, such as "low, middle and high", are referred to as continuous categories.

Property

A quantitative parameter characteristic of the objects for which a functional representation is desired (dependent variable).

Training Set

The list of data vectors used to generate classification or prediction rules.

Evaluation Set

The list of data vectors used to evaluate the performance of classification or prediction rules.

Test Set

The list of data vectors, in classification or prediction analysis, for which the true category or property value is not known. The Evaluation and Test sets are functionally one and the same. The Evaluation set is a "let's pretend" Test set.

Uncertainty

The error associated with an analytical measurement. The uncertainty is assumed to include all sources of errors such as sampling, instrumental, chemical, etc.

It should be recognized that these definitions are not particularly rigid or mutually exclusive. A continuous property can certainly be segmented into the low resolution categories "too low" and "high enough". The parameter considered as a property in one phase of analysis may well be a measurement in another. The Training and Evaluation set definitions may be switched. It may even be desired to switch the definition of object and feature. If the data are considered as a matrix (objects as rows and features as columns), the switch is equivalent to the transposition of the matrix. And it is certainly good practice, no matter what the specific nature of the data analysis problem, to make at least cursory unsupervised data analysis, if nothing more than to give a rough screen for some gross, unsuspected structure in the data.

Pattern Recognition: New Techniques that Utilize Analytical Error

The general problem that is amenable to solution by the techniques available in ARTHUR is the analysis of patterns in an n-dimensional space. In the past, applications utilizing pattern recognition have not taken into account the errors in the measurements because the mathematical methods currently available make no provision for their inclusion. However, in most chemical data the inadvertent assignment of zero measurement error which results is clearly an unrealistic assumption. This problem has been investigated by Fasching, Duewer and Kowalski (3). As a result of this study, several algorithms in ARTHUR have been modified to include the uncertainties in the calculations.

Current pattern recognition techniques treat measurements as dimensions in an n-dimensional space. If, for each member of a collection of objects (samples), n measurements are known, the samples are represented as points in the space formed by the measurements. Therefore, the value of a given measurement for a particular object serves to exactly position the point representing the object along a coordinate measurement axis in the n-space. Figure 1 depicts the configuration of the data vectors from two samples in a three-dimensional space. The set of all such vectors defines the data matrix.

In analytical applications, where the uncertainties in the measurement are either known or can be estimated, there exists a matrix of uncertainties corresponding to and of the same dimensions as the data matrix. Mathematical operations that transform the data matrix also change the uncertainty matrix to a transformed uncertainty matrix. Each element of the uncertainty matrix reflects the exactness (in units of $\pm$ one standard deviation) to which the corresponding element of the data matrix is known. Therefore, each measure-

ment in the data matrix is now treated as a mean value with a probability distribution defined by its error. The effect of the inclusion of uncertainties on the vectors in Figure 1 is illustrated in Figure 2. The analytical uncertainties reflect the fact that a data vector is not, in reality, a point in the measurement space, but is the most probable value in a region of probability in this space. If the area of the elipsoid in this example is defined at a 50% probability level of the standard deviation of each measurement, then another set of measurements made on a sample would have an equal probability of lying outside the elipsoid as within it. This model is more reasonable for most chemical problems.

At present, ARTHUR has been modified to include the analytical error in representative method of preprocessing, display, supervised learning and unsupervised learning. A full description of these modifications can be found in reference 3. The current methods deal only with symmetric uncertainties. A nonmetric (unsymmetrical) distance is defined; however, classification and clustering routines utilizing distance have not, as yet, been similarly modified to make use of this type of distance matrix.

Since the uncertainty matrix contains information about the error associated with each measurement, it can be incorporated into the preprocessing of the data matrix. The more realistic features generated can be utilized in all reported methods of pattern recognition, thus eliminating the need to change each analysis method. The scaling algorithms (SCALE)* in ARTHUR have been modified to include uncertainties. An error-weighted mean and variance are utilized in place of the feature mean and variance in these calculations. The new mean of the j^{th} feature in the data is defined as:

$$\bar{x}_j = \frac{\sum_{i=1}^{m} x_{i,j}/u_{i,j}^2}{\sum_{i=1}^{m} 1/u_{i,j}^2}$$

*Methods (names in capital letters) are described in appendix

where the $u_{i,j}$'s are the entries in the uncertainty matrix corresponding to the data matrix measurement $x_{i,j}$, and the sum is over the training set data vectors. Modification of the available distance metrics have also been made along with the addition of new distance calculations based on measurement uncertainties. The algorithms for these can be found in the appendix (DISTANCE) to this chapter. The modified city-block distance and the modified Mahalanobis distance are now weighted by a function of the measurement errors associated with the features going into the calculation. A new metric, the gaussian overlap-integral distance, greatly emphasizes the features that have a small distribution with respect to their measurement size and related uncertainties. A maximum distance of one is assigned to features that differ greatly from each other or have very small uncertainties. Another new distance calculation, the gaussian feature-space distance, calculates a distance value that is proportional to the probability that a feature in the i^{th} data vector belongs to the same population as the corresponding feature of the j^{th} data vector. These are summed over all the feature space to give the intersample distance. The calculation is nonmetric and the distance matrix is unsymmetrical.

The uncertainty matrix has also been incorporated in the Karhunen-Loève transform. The modified technique transforms the uncertainties into a new certainty matrix along with the sample matrix. The assumption is made that the same degree of correlation applies to the uncertainty matrix as is used in the transformation of the sample matrix.

Introduction to Data Analysis Using ARTHUR

Different collections of objects may have quite different data structures varying from a random scatter to well defined clusters or curvilinear shapes. Since each algorithm affects data reduction according to the criterion upon which it is based, a thorough understanding of the inherent assumptions imposed upon the data structure in the formulation of a technique and the limitations that may result can provide information helpful in arriving at an understanding of the underlying structure of the data when the methods are applied in combination.

Suppose, for example, the n-dimensional structure of two categories of objects we wish to separate by pattern recognition classification techniques corresponds to the one dimensional problem depicted in Figure 1, where the shaded portions of the figure correspond to category 1 and the unshaded portions to category 2.

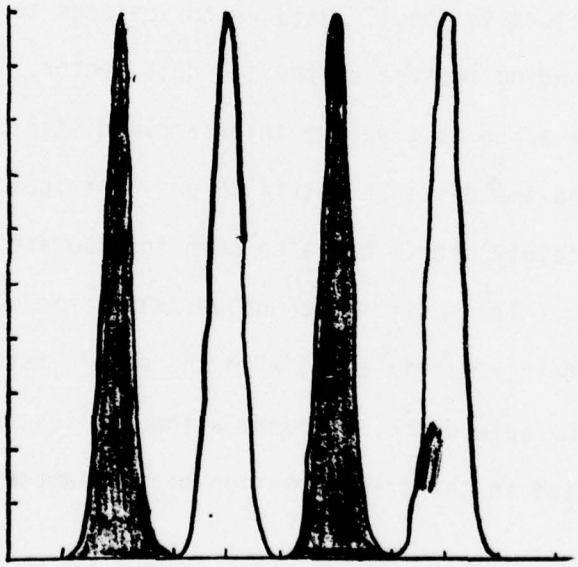


Figure 1. Bimodal distribution

Whereas in one dimension the solution to the problem is obvious, in n-dimensions the bimodality may not readily reveal itself. If PLANE or SIMCA were applied to these data, the results might lead one to believe

that the categories cannot be distinguished since the data are neither linearly separable nor continuous. However, KNN would encounter no problems since the objects in the near vicinity of a given point tend to be of its class. Bayesian classification (as long as no a priori distribution is assumed) would also produce good results. (Note that plots of the data might expose this distribution in a less ambiguous form. Consequently, this example is meant only as an illustration of the effects of the methods on an easy to understand distribution.)

Unfortunately, the solution to a real problem does not, in general, tend to be as straightforward and may require a great deal of interaction and guidance from the analyst aided by pre-processors, display methods, and statistics. For this reason, the capabilities of ARTHUR for displaying the data are quite well developed when combined with the ingenuity of the analyst as will be seen in a later section of this chapter. On the other hand since preprocessing refers to any method that translates, rotates, or in any way transforms the data, such infinite diverse possibilities arise that were we to include only those methods that we and others have found useful, they would dominate the code. Therefore, the set of preprocessing tools available in ARTHUR is aimed mainly toward normalization, feature weighting, and dimensionality reduction. In addition, ratios of features can be added to the feature list in TUNE and individual features can be transformed or combined in CHFEATURE. Since the methods chosen to preprocess the data can ultimately determine success or failure in the solution of a problem by pattern recognition methods and/or the cost of the analysis, methods not available in ARTHUR should not be neglected. An example of this is the utilization of the Fourier, Hadamard and autocorrelation functions for the transformation of spectral data (6, 7).

Two assumptions made throughout any supervised pattern recognition

technique are that the features used contain information useful to the solution of the problem and that, even when this is known to be the case, the data can be transformed into a representation amenable to the algorithms employed. When this is not the case it may become necessary to either change the form of the question being asked about the existing data or redesign the experiment from which the measurements are obtained. Hopefully, the information gained through prior analysis will serve to guide the analyst in this endeavor.

We have discovered that techniques originally designed for unsupervised learning applications are powerful tools in the early stage of all data analysis problems. These methods have seen little application in chemistry. Since the goal of these methods is the determination of the existence of inherent data structures within a larger data structure, neither training nor classification is attempted. TREE and HIER are two unsupervised learning methods which are based on the similarity of objects as defined by their distances in the feature space. Factor analysis can also be utilized in this mode.

The following sections are a brief description of the basis of the various pattern recognition methods used to analyze this data:
WEIGHT is a preprocessing method that weights each feature on the basis of its individual importance to the solution of a pattern recognition problem. For categorized data, the criterion of importance can be either the total variance or total Fisher weight for the feature. The variance weight is a ratio of the interclass variance of two categories to the intraclass variances of the categories. If $W_{j,m,n}$ is a measure of the utility of feature j in separating categories m and n, the variance weight $(WV)_{j,m,n}$ is:

$$(WV)_{j,m,n} = \frac{\sum_{k=1}^{Nm} \frac{x_{k,m,j}^2}{Nm} + \sum_{k=1}^{Nn} \frac{x_{k,n,j}^2}{Nn} - \frac{2\sum_{k=1}^{Nm} x_{k,m,j} \sum_{k=1}^{Nn} x_{k,n,j}}{NmNn}}{\sum_{k=1}^{Nm} \frac{(x_{k,m,j} - \bar{x}_{m,j})^2}{Nm} + \sum_{k=1}^{Nn} \frac{(x_{k,n,j} - \bar{x}_{n,j})^2}{Nn}}$$

where N_i is the number of data vectors in category i; the total variance weight is the geometric mean of the individual category pair weights. The Fisher weight is a ratio between the square difference in the category pair means and the sum of intraclass variances:

$$(WF)_{j,m,n} = \frac{(\bar{x}_{m,j} - \bar{x}_{n,j})^2}{\sum_{k=1}^{Nm} \frac{(x_{k,m,j} - \bar{x}_{m,j})^2}{Nm} + \sum_{k=1}^{Nn} \frac{(x_{k,n,j} - x_{n,j})^2}{Nn}}$$

The total Fisher weight is the arithmetic average of the individual category pair weights.

For continuous property data the weighting is done on the basis of the correlation of the feature to the property. The square correlation to property of feature j is:

$$\frac{\left[\sum_{k=1}^{N} (x_{j,k} - \bar{x}_j)(p_k - \bar{p})\right]^2}{\sum_{k=1}^{N} (x_{j,k} - \bar{x}_j)^2 \sum_{k=1}^{N} (p_k - \bar{p})^2}$$

where N is the number of data vectors in the training set and p_k is the property of the k^{th} data vector.

SELECT (28) is a feature selection technique that generates orthogonal features based on their importance to classification. The criterion for importance for categorized data is the variance or Fisher weight and for continuous-property data, the correlation-to-property weight (see WEIGHT). The highest weighted feature is selected as the first feature. The remaining features are then decorrelated from the chosen feature. The de-

correlated features are reweighted and the feature whose new weight is highest becomes the second selected feature. The process continues until either a specified number of features is chosen or a given minimum weight attained. The selected (unweighted) features are output to a file for later use. The user can opt for the decorrelated features or the same features in their unchanged form. Since one set is a linear combination of the other set, the same information is retained for either option. Only the representation is changed (i.e. the sub-feature space is either rotated or not rotated to orthogonal axes).

GRAB. As a feature selection method, GRAB (12) is intermediate between weight (with no feature decorrelation) and the more expensive SELECT (with total decorrelation). A previously-weighted file of n data vectors is input to the routine. Each feature is assigned an initial weight

$$W(1)_i = \{\sum_{k=1}^{n} (x_{i,k}-x_i)^2\}^{\frac{1}{2}}$$

The feature with the largest weight is selected as the first new feature. Each of the remaining features is reweighted such that if $C_{i,j}$ is the correlation between the i^{th} feature just chosen and the remaining feature j,

$$W(2)_j = W(1)_j[1-|C_{i,j}|]$$

For the m^{th} iteration the weight of the j^{th} feature remaining is

$$W(m)_j = W(1)_j \prod_{i=1}^{m-1} [1-|C_{i,j}|]$$

LEAST performs a least-squares multi-linear regression that is best suited to continuous property problems. If D is a data matrix with associated property matrix p, then $W=(D^TD)^{-1}D^TP$ is the least squares solution to the set of linear equations P=DW where W is a vector which weights the utility of the features in fitting the data.

In actual practice, determination of the weight vector is done by

$$W = [E^T C^{-1} E] X^T P$$

where X is obtained by mean normalization of D, C^{-1} is the inverted correlation matrix associated with D and E is a diagonal matrix whose elements are the reciprocal variances of the features.

Prediction of an unknown property P' is based on the weight vector obtained is therefore

$$P' = X'W$$

LEDISC is a multi-linear least squares regression designed for categorized data. Except in property definitions it is computationally equivalent to LEAST. For a data set of n categories, n linear regressions are performed such that for the i^{th} regression the property P is defined as

$$P = \begin{cases} +1 \text{ for all vectors in category } i \\ 0 \text{ for all vectors not in category } i \end{cases}$$

An unknown data vector is placed into that class whose weight vector produces the largest value.

LESLT is a variable reduction technique which seeks to optimize category pair separation in as few variables as possible. A feature derived is a linear combination of the original data that describe the position of a data vector relative to a hyperplane between two categories in the data set. The input data matrix (X) of n categories is divided into $n(n-1)\frac{1}{2}$ submatrices. If Y is the submatrix containing only those patterns in categories i and j plus the test data, an outcome column matrix of properties can be defined such that

$$G^{i,j} = \begin{cases} -1 \text{ for patterns in } i \\ +1 \text{ for patterns in } j \end{cases}$$

Thus defined, there exists a vector W_k of weights such that $YW_k = G^{i,j}$. (Determination of W_k is the least squares solution for this equation (see LEAST).) The weight vector obtained is used to transform and classify

all the data vectors in Y. This process is followed for all category pairs. Once all the weight vectors are obtained, the entire data matrix (X) is transformed such that X' = XW. The new matrix X' obtained has n(n-1)/2 features which are approximate category-pair separators.

LEPIECE does a piece-wise least squares multiple regression for each data vector in the training and test set. The property of each data vector is predicted from the fit (see LEAST) using the k-nearest-neighbors (see KNN) to the vectors. The value of k is a user-defined multiple of the number of features. The criterion used for "nearest" is the inter-pattern distance (see DISTANCE). Only those features used in the determination of the distance are used in the regression.

MULTI is a hyperplane discriminant function method designed for multi-category data. Computationally, it is equivalent to PLANE, except in category definition. For a data matrix of n categories, n hyperplanes are generated such that the i^{th} hyperplane describes the separation of the i^{th} category from the rest of the data.

PLANE generates and classifies on the basis of a linear discriminant function and is best suited to data containing two categories (see MULTI for multicategory case). By an error-correction feedback method it seeks a hyperplane in an augmented n+1 space (where n is the number of features) that best separates a pair of categories.

Each data vector in n space is considered a vector in n+1 space where the $n+1^{th}$ feature is unity. Therefore, two classes can be defined as lying on either side of a hyperplane (whose equation in n+1 space is $W \cdot Y = \emptyset$), through the origin with corresponding class numbers +1 and -1. The discriminant function is calculated by first loading a weight vector with random or user-defined values. During training, classification of vector Y_k by this weight vector is a decision of the form

$$W \cdot Y_k = S_k = \begin{cases} \underline{\text{correct}}, \text{ if the sign of the response relative to the hyperplane is the same as the sign of its class} \\ \underline{\text{incorrect}}, \text{ if the sign is not the same} \end{cases}$$

If a pattern is misclassified, the weight vector is adjusted by reflection of the hyperplane about the misclassified point. The new weight vector is then used to classify the data. The process continues until all patterns in the training set are correctly classified or a maximum number of iterations are reached.

For more than two categories, a hyperplane separating each pair of categories is found. An unknown data vector is then classified using a majority committee vote procedure on all the discriminant function responses. The use of PLANE for multi-category data is equivalent to a piece-wise learning machine.

REGRESS is a multidimensional multivariate regression method which computes a linear discriminant function. It accepts both category and continuous data. Two optimization methods are available. Either the residual variance or the multiple correlation can be minimized.

STEP is a stepwise multi-linear regression method. Features used in the regression are determined by their contribution to the overall variance. In the regression, features are added one at a time such that the feature that is added makes the greatest improvement in the "goodness of fit." When a feature that is indicated to be significant to the reduction in variance in an early stage of the regression is indicated to be insignificant after the addition of several other features, it is eliminated from the regression before addition of another feature. The criterion for selection of a feature to add or remove from the calculation is as follows:

Removal: If the variance contribution is insignificant at a specified F-level, the feature is removed from the regression.

Addition: If the variance reduction due to addition of a feature is significant F-level, this feature is entered into the regression.

HIER is an unsupervised learning (cluster analysis) method based on the relative similarity of a set of data vectors. Each vector is initially assumed to be a lone cluster. A similarity matrix is constructed such that if $S_{i,j}$ is the similarity between the i^{th} and j^{th} data vector, then

$$S_{i,j} = 1 - \frac{d_{i,j}}{d_{max}}$$

where $\frac{d_{i,j}}{d_{max}}$ is the interpattern distance of data vectors "i" and "j" normalized by the largest interpattern distance d_{max} in the data (see DISTANCE).

The matrix is scanned for the maximum similarity in the set. These "most similar" vectors are clustered, removed from the matrix and replaced by a new vector whose location is the average of the two vectors. In combining clusters, two options are available. Either the average of the two clusters is weighted by the number of data vectors in each cluster or each cluster is given equal weight. The new matrix is scanned for the next greatest similarity and the procedure is repeated. The process ends when all the data vectors form a single cluster. Output is in the form of a connection dendrogram.

Literature Cited

1. Duewer, D. L., Harper, A. M., Koskinen, J. R., Fasching, J. L. and Kowalski, B. R., ARTHUR, Version 3-7-77.

2. Koskinen, J. R. and Kowalski, B. R., Journal of Chemical Information and Computer Science, 15, 119 (1975).

3. Fasching, J. L., Duewer, D. L. and Kowalski, B. R., submitted to Analytical Chemistry.

4. Nie, N. H. et al., "Statistical Package for the Social Sciences, 2nd Edition" McGraw Hill, Inc., New York, 1975.

5. Dixon, W. J., Ed., "BMD, Biomedical Computer Programs", University of California Press, Berkeley, 1971.

6. Kowalski, B. R. and Reilley, C. A., Analytical Chemistry, 42, 1387 (1971).

7. Kowalski, B. R. and Bender, C. F., Analytical Chemistry, 45, 2334 (1973).

8. Miller, R. G., Biometrika, 61, 1 (1974).

9. Duewer, D. L., et al., Analytical Chemistry, 48, 2002 (1976).

10. Kowalski, B. R. et al., Analytical Chemistry 44, 2176 (1972).

11. Stevenson, D. F. et al., Archaeometry (1971) 13, 17.

12. Duewer, D. L. et al., "Documentation for ARTHUR, Version 1-8-75", (1975) Chemometrics Society Report No. 2.

13. Reinsch, C. H. Numerische Mathematik (1967) 10, 177.

14. Birnbaum, Z. W., JASA (1952) 47, 425.

15. Davies, O. L. and Goldsmith, P. L., "Statistical Methods for Research and Production", p. 234, Hafner, New York, 1972.

16. Anderson, T. W., "An Introduction to Multivariate Statistical Analysis", p. 65, John Wiley & Sons, New York, 1958.

17. Mahalanobis, P. C. Proceedings of the National Institute of Science in India, p. 49, 122, (1936).

18. Anders, O. U., Analytical Chemistry, 44, 1930 (1972).

19. Any numerical analysis text.

20. Horst, P., "Factor Analysis of Data Matrices", Holt, Rinehart and Winston, Inc., New York, 1965.

21. Wold, S., Journal of Pattern Recognition, 8, 127 (1976).

22. Kowalski, B. R. in "Computers in Chemical and Biochemical Research, Vol. 2", Academic Press, New York, 1974.

23. Andrews, H. C., "Introduction to Mathematical Techniques in Pattern Recognition", Wiley Interscience, New York, 1972.

24. Kowalski, B. R. and Bender, C. F., Journal of the American Chemical Society, 95, 686 (1973).

25. Cover, T. M. and Hart, P. E., IEEE Transaction of Information Theory, IT-13, 21 (1967).

26. Kowalski, B. R. and Bender, C. F., Journal of the American Chemical Society, 94, 5632 (1972).

27. Fisher, R. A., Annals of Eugenics, 7, 179 (1936).

28. Kowalski, B. R. and Bender, C. F., Journal of Pattern Recognition, 8, 1 (1976).

29. Kowalski, B. R., Analytical Chemistry, 41, 695 (1969).

30. Kowalski, B. R. and Bender, C. F., Analytical Chemistry, 45, 590 (1973).

31. Nilsson, N. J., "Learning Machine", McGraw Hill, New York, 1965.

32. Ralston, A., and Wolf, H. S., "Mathematical Methods for Digital Computers", p. 191, John Wiley and Sons, New York, 1966.

33. Hamming, R. W., "Introduction to Applied Numerical Analysis", McGraw Hill, New York, 1971.